JRC Scientific and Technical Reports

Uncertainty assessment for PA models

Ricardo Bolado-Lavin, Anca Costescu Badea

JRC EUROPEAN COMMISSION

Institute for Energy

EUR 23503 EN - 2008

The Institute for Energy provides scientific and technical support for the conception, development, implementation and monitoring of community policies related to energy. Special emphasis is given to the security of energy supply and to sustainable and safe energy production.

European Commission
Joint Research Centre
Institute for Energy (IE)

Contact information
Address: P.O. Box 2, 1755ZG Petten, The Netherlands
E-mail: Anca.COSTESCU-BADEA@ec.europa.eu
ricardo.bolado-lavin@jrc.nl
Tel.: +31 224 565131
Fax: +31 224 565641

http://ie.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC47247

EUR 23503 EN
ISBN 978-92-79-09779-9
ISSN 1018-5593
DOI 10.2790/17625

Luxembourg: Office for Official Publications of the European Communities

Printed in The Nederland

0. Abstract

A mathematical model comprises input variables, output variables and equations relating these quantities. The input variables may vary within some ranges, reflecting either our incomplete knowledge about them (epistemic uncertainty) or their intrinsic variability (aleatory uncertainty). Moreover when solving numerically the equations of the model, numerical errors are also arising. The effects of such errors and variations of the inputs have to be quantified in order to asses the model's range of validity. The goal of uncertainty analysis is to asses the effects of parameter uncertainties on the uncertainties in computed results.

The purpose of this report is to give an overview of the most useful probabilistic and statistic techniques and methods to characterize uncertainty propagation. Some examples of application of these techniques for PA applied to radioactive waste disposal are given.

1. Notation

In the following, the random variables (or *variates*) will be denoted by upper-case letters, while their realizations will be denoted by the corresponding lowercase letters. The letter X (x) will be associated with the input parameters and the letter Y (y) with the output.

rv : random variable
X, Y : random variables;
$(X_1, X_2, ..., X_n)$: a random sample ;
$(x_1, x_2, ..., x_n)$: the corresponding realization of the random sample;
$\boldsymbol{X} = (X_1, \ldots, X_d)$: a random vector of size d ;
$Y=Y(\boldsymbol{X})$: the output of the numerical model ;
F : the cumulative distribution function (CDF): $F(x) = P(X \le x)$;
f : the probability density function (PDF) : $F(x) = \int_{-\infty}^{x} f(t)dt$;
IR: set of real numbers;
x_α, q_α : the α- quantile of X, defined as $F(x_\alpha) = \alpha$;
$[x]$: the largest integer $\le x$;
$\lceil x \rceil$: the smallest integer $\ge x$;
$X_{(k)}$: order statistics (of order k) ;
μ : mean of a random variable;
σ^2 : variance of a random variable;
$\bar{x}$: sample mean;
σ_x^2, s^2 : sample variance; σ_x, s : sample standard deviation
E(.) : mathematical expectation
Var(.) : variance
iid : independent, identically distributed

2. Introduction

A performance assessment (PA) of a repository for radioactive waste is an analysis that identifies the processes and events that might affect the disposal system; examines the effects of these processes and events on the performance of the disposal system; and, estimates the cumulative releases of

radionuclides, considering the associated uncertainties, caused by all significant processes and events. These estimates shall be incorporated into an overall probability distribution of cumulative release to the extent practicable.

The process to develop a Performance Assessment of a nuclear High Level Waste repository (HLW) involves modelling the whole system, which classically is considered to be divided into three parts: i) The near field or engineered facilities including the disturbed part of the geosphere, ii) the far field or part of the geosphere that hosts the repository, and iii) the biosphere, eventual sink of radioactive pollutants. Modelling such a system means modelling the inventory of radionuclides, the processes that deteriorate the facility and that produce the release of radionuclides in the long term, their transport through the geosphere and their spread over the biosphere, which ultimately will produce doses on humans. All those models will be integrated as submodels of the system model.

Moreover, implementers should be able to foresee potential disruptive scenarios that could induce 'worse than expected' behaviour of the system. This involves addressing events and processes that, though unlikely, could reasonably happen, and would produce more adverse consequences that the expected normal evolution of the system. Two activities are triggered when alternative scenarios are identified: Likelihood estimation and adapting the system model to the specific physical and chemical conditions produced by the scenario.

Parameters such as coefficients, boundary and initial conditions of the differential equations used in the system model are usually affected by uncertainty. Characterising these uncertainties is an extremely time consuming task that includes laboratory and field experiments, collection of historical records, search in databases and use of expert judgment. Formally, as soon as scenarios are identified and their probabilities are estimated, the system model is available and parameter uncertainty is assessed, computations could be started to estimate the adverse consequences to humans and the environment in the future.

This report focuses on three steps of the PA:

- the characterisation of input parameter uncertainties and scenario likelihood,
- the propagation of uncertainties and
- the characterisation of output uncertainties (see Figure 1).

Regarding the characterisation of input parameter uncertainties and scenario likelihood estimation, special attention will be paid to widely used descriptive statistics (chapter 3), which are useful to understand the data obtained from field and laboratory experiments, and to inference methods used for assigning PDFs to rvs and probabilities to scenarios (chapter 4). Chapter 5 deals with the propagation of uncertainties through the system model, and specifically with the most useful, and in fact most used, method: Monte Carlo. Further attention is paid within this chapter to techniques designed to make Monte Carlo computationally more efficient (variance reduction techniques and input parameter space dimension reduction) and to the use of surrogate models. Some pages are also dedicated to the selection of the sample size and other uses of Wilk's theory on tolerance intervals. The last chapter is dedicated to specific issues related to the characterisation of output uncertainties.

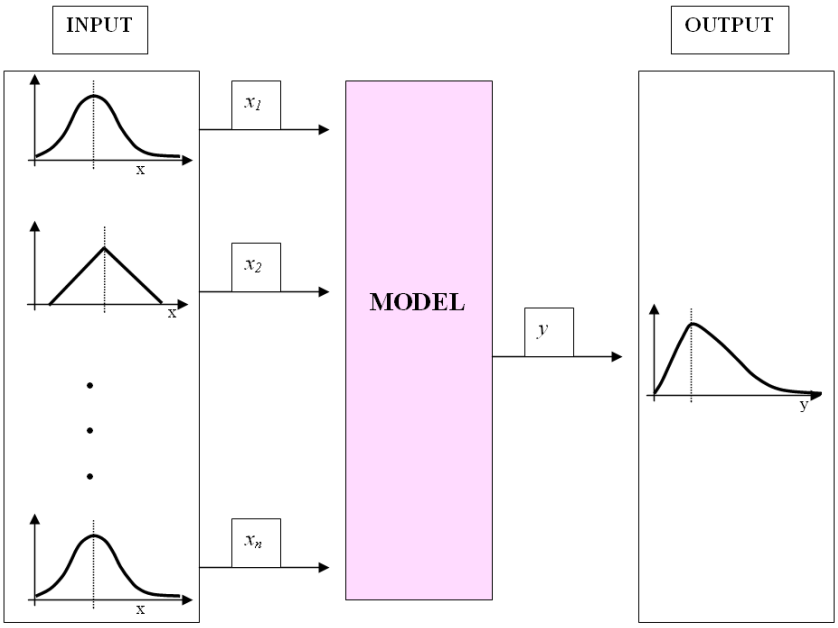


Figure 1: Uncertainty propagation using a single output model

3. Descriptive statistics

Descriptive statistics are used to get information about the data that have to be analysed. The techniques described in this chapter are generic statistical techniques that can be used to analyse data related to the inputs, but also to the outputs. [Prváková 08] has been used as a source of data to be used in some of the examples shown in this report. The realizations of the rv are generically denoted by $(x_1, x_2, ..., x_n)$.

In the following the word **statistic** will design "a function of a sample where the function itself is independent of the sample's distribution: the term is used both for the function and for the value of the function on a given sample" (from http://en.wikipedia.org/wiki/Statistic).

3.1. Numerical summaries

3.1.1. Central tendency

The purpose of the measures of central tendency, or location, is to compute one single number which gives the best possible representation of the value around which the data are located. Four measures have been considered: mean, median, geometric mean and mode.

The mean

The most important one is the arithmetic **mean** of the sample, defined by:

$$\bar{x} = 1/n \sum_{i=1}^{n} x_i \,. \qquad (3.1)$$

Other notations: $\bar{x}_n$ (whenever knowing the sample size is needed), μ.

Important characteristics of the arithmetic mean:

- It is a linear statistic in the following sense: for two samples x and y of the same size $\overline{ax+by} = a\bar{x} + b\bar{y}$ ($a,b \in IR$).
- It is not a robust statistic; it is very sensitive to extreme values (see the example in next page).
- It gives a very good measure of location for homogeneous symmetric sets of data. The variance (see § 3.1.3) is minimised when mean is the measure of location used as a reference.
- If the sample is heterogeneous (existence of data obtained under different conditions), the mean can become completely useless as a measure of central tendency (this problem affects to all measures of central tendency); it could even take a value outside the range of definition of the variable under study. For instance a sample made of two subsamples of equal size that do not overlap at all, the arithmetic mean would be in between, just where the variable takes no value.
- When many output variables are used in a PA, the values may spread over several orders of magnitude, and the aggregated may become very asymmetric and completely dominated by the largest sample values. In such case the arithmetic mean is generally not a good statistic.

The geometric mean

The geometric **mean** of the sample is defined by:

$$\tilde{x} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}. \tag{3.2}$$

The geometric mean is only of interest when all sampled values are positive. It may also be computed when there are null values, but then it is also null. The geometric mean gives a measure of central tendency when a logarithmic scale is used. The geometric mean is often approximated by calculating the arithmetic mean of the logarithm of the actual sampled values and transforming the obtained result consequently, i.e.:

$$\tilde{x} = \exp\left[(1/n)\sum_{i=1}^{i=n} \ln(x_i)\right] . \tag{3.3}$$

The geometric mean is always either equal to or smaller than the arithmetic mean. In cases when positive and null values are mixed in the same sample, it may be of interest to compute a geometric mean restricted to the m sampled positive values (m<n). The geometric mean does also answer the question "if all quantities had the same value, what would that value have to be in order to achieve the same product".

As many rv used as outputs in PA studies are spread over several orders of magnitude, the geometric mean is useful to estimate the "center of mass" of the data in a logarithmic scale (the arithmetic mean estimates the "center of mass" in a linear scale).

The median

Another alternative is the **median.** The median of a sample is the value which splits the sample in two equal parts. Suppose that the sample sorted in ascending order is $x_{(1)}, x_{(2)}, ..., x_{(n)}$, then

$$med(x) = \begin{cases} (x_{(n/2)} + x_{(n/2)+1})/2 & \text{if } n \text{ is even} \\ x_{(n+1)/2} & \text{if } n \text{ is odd} \end{cases} . \tag{3.4}$$

It is the value of x for which the CDF $F(x) = 1/2$.

The median is a robust indicator, but it is more difficult to perform algebraic computations using it than using the mean. For instance, the linearity property is no longer valid. On the other hand, the median is conserved when applying a strictly monotonic increasing transform to the sample, which is not the case for the mean.

Example:

The sample data (n=51) represents the release of ^{94}Nb getting out of the fractured zone after 5000 years computed in a study of the release of radio nuclides from wastes of an ILW disposal cell embedded in a porous material for a generic French clay site (see [Prváková 08] for the description of the benchmark). In Figure 2 the circled point contains an extreme value.

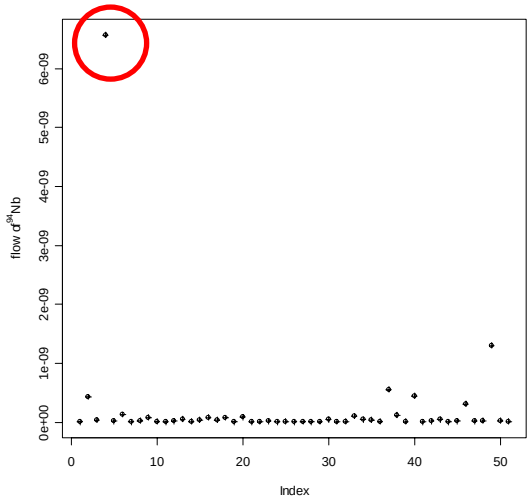


Figure 2: A sample of ^{94}Nb getting out of the fractured zone after 5000 years

The mean of the whole sample is $\bar{x}_{51} = 2.18e-10$, while if we exclude the extreme point we obtain $\bar{x}_{50} = 9.13e-11$. The effect on the mean of that single value is huge; excluding it from the sample produces a decrease of 58% in the mean. This is not the case with the median. The original median (sample of size 51) was $med_{51}(x) = 2.22e-11$. After removing the extreme value, the new median is $med_{50}(x) = 2.21e-11$; the two values are quite similar, which is due to its robustness as a measure for the central tendency.

The mode

A mode is the location of a local maximum of the PDF. A PDF can be multimodal, which often means that we are dealing with heterogeneous populations. For discrete data, the mode is the most frequently observed value. However, the estimation of the mode using a sample depends entirely on the method used to estimate the PDF (see section 3.2.3).

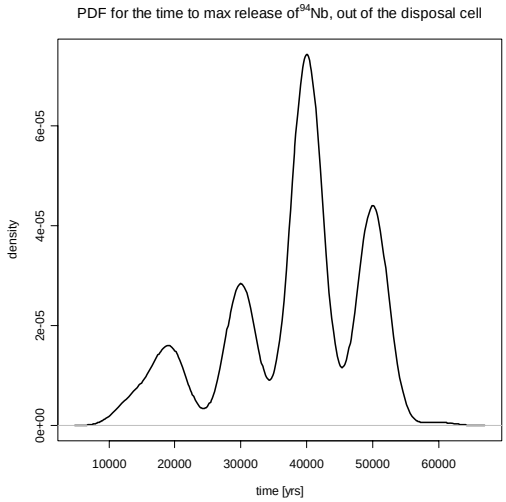


Figure 3: Example of a multimodal PDF; there are 4 modes: 19000, 30000, 40000 and 50000 years

3.1.2. Quantiles

Quantiles generalize the median for a probability α different from ½, i.e. they are values that split the data in two parts, such as the proportion of data inferior or equal to this value is equal to α. The α-quantile q_α is defined by the equation

$$F(q_\alpha) = \alpha\ ,\ \forall \alpha \in [0,1]\,. \tag{3.5}$$

However, when the cumulative distribution is not strictly increasing function this equation might have either an infinite number of solutions or no solution at all, as can be seen from Figure 4. The usual conventions to overcome this problem are based on the ordered observations $x_{(1)} \le \ldots \le x_{(n)}$. The smallest observation corresponds to a probability of 0 and the largest one to a probability of 1. The i^{th} observation corresponds to α-quantile q_α (i.e. $q_\alpha = x_{(i)}$), where α may be defined as follows:

$$\alpha = \begin{cases} \boldsymbol{(i-1)/(n-1)} & \text{or} \\ (i-0.5)/n & \text{or} \\ i/(n+1) & \text{or} \\ \boldsymbol{i/n} & \text{or} \\ (i-1/3)/(n+1/3) & \text{or} \\ (i-3/8)/(n+1/4). \end{cases} \tag{3.6}$$

In (3.6), the two emphasized expressions are the most used ones:

- the first one because it has a symmetry with respect to the CDF: the smallest observation corresponds to a probability of 0 and the largest one to a probability of 1; it is the one used by default by some statistical softwares such as R [http://cran.r-project.org/] and S [http://www.insightful.com/];
- the fourth one because it corresponds exactly to the empirical cumulative distribution function (see formula (3.13)).

Concerning the other expressions in (3.6):

- the second one is popular amongst hydrologists,

- the third one is used by other statistical softwares such as Minitab [http://www.minitab.com/] and SPSS [http://www.spss.com/],
- using the fifth expression, one obtains a quantile estimate that is approximately median-unbiased (i.e. the median of the estimator is approximately unbiased) regardless of the distribution of *x*,
- using the last expression, one obtains a quantile estimate that is approximately unbiased for the expected order statistics if *x* is normally distributed.

More details should be found in [Hyndman 96].

If α is not of one of the previous forms, a linear interpolation may be used to estimate q_α, as for example $\alpha = (1-a)\frac{(i-1)}{n-1} + a\frac{i}{n-1} \Rightarrow q_\alpha = (1-a)x_{(i)} + ax_{(i+1)}$, $0 < a < 1$.

Example:

Let us consider the following *n*=6 sample $\{0,1,2,3,4,5\}$. We assume that the i^{th} observation is the estimation of the α-quantile q_α, where $\alpha = (i-1)/(n-1)$. We want to estimate the 1/4 quantile, which is not of the form $\alpha = (i-1)/(n-1)$.

But, as $\frac{1}{5} < \frac{1}{4} < \frac{2}{5}$ there exists $a = 0.25$, $(0 < a < 1)$, and $i = 2$, such that $\frac{1}{4} = (1-a)\frac{i-1}{5} + a\frac{i}{5}$. We hence obtain $q_{1/4} = 0.75x_{(2)} + 0.25x_{(3)} = 0.75 \times 1 + 0.25 \times 2 = 1.25$.

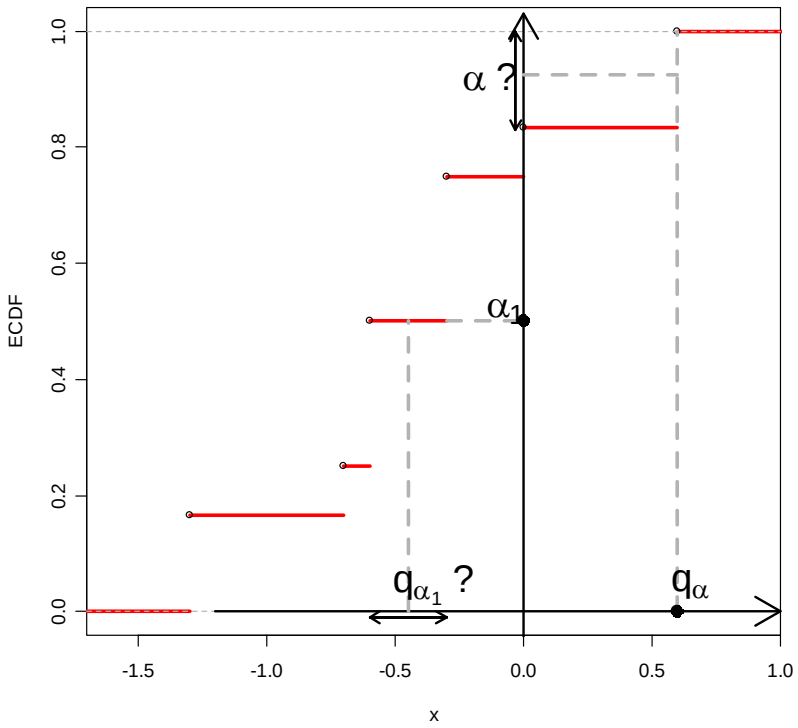


Figure 4: Empirical cumulative distribution function and quantile signification

The median is the ½ quantile. Some other particular quantiles frequently used are:

- percentiles, the 1/100-quantiles
- deciles, the 1/10-quantiles
- quartiles, the 1/4-quantiles.

For more details concerning quantile estimation see section 5.4.

3.1.3. Dispersion characteristics

The measures of dispersion are important for describing the spread of the data around a central value. Two distinct samples could have similar means or medians but completely different degrees of dispersion around them.

The range

The range is defined as the difference between the largest and smallest sample values:

$$range = x_{(n)} - x_{(1)} = \max(x) - \min(x)\,. \tag{3.7}$$

It is one of the simplest measures of variability to calculate, but it depends only on extreme values (and hence it is a non robust indicator) and provides no information on the data distribution.

The interquartile range (interval)

The interquartile range is defined as the difference between the 3rd and the 1st quartiles, i.e. $q_{3/4} - q_{1/4}$. It is a robust indicator. The meaning of this indicator is that at least 50% of the "central" data are contained in this interval. It is also used for drawing the boxplots (see section 3.2.4).

The variance and the standard deviation

This indicator was meant to measure the mean deviation from the mean value of the sample, by taking into account positive and negative deviations. This is the reason for introducing the quadratic function **sample variance** as:

$$\mathrm{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\ . \tag{3.8}$$

As the variance does not have the same units as the sample (because of the squares), the **standard deviation** has been introduced:

$$\sigma_x = \sqrt{\mathrm{var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ . \tag{3.9}$$

Alternative definitions of the variance and the associated standard deviation are

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\ ;\ s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.10}$$

(see for instance [Saporta 90] for more information about the different definitions of the variance).

If the sample is approximately normal, then

1. The interval mean ± one standard deviation contains approximately 68% of the measurements in the series.
2. The interval mean ± two standard deviations contains approximately 95% of the measurements in the series.

3. The interval mean ± three standard deviations contains approximately 99.7% of the measurements in the series.

When the distribution that generates the sample is unknown, similar rules, based on Chebyshev's inequality [Jordaan 05], may be applied. However, the bounds that are computed are rather loose, but they are valid irrespective of the distribution that generates the data; knowing mean and standard deviation is enough to calculate them.

The Chebyshev's inequality states that $fr(|x_i - \bar{x}| \geq ks) \leq 1 - 1/k^2$, where k is any real number and fr stands for relative frequency. It provides useful information for $k>1$:

1. The interval mean ± two standard deviations contains at least 75% of the measurements in the series.
2. The interval mean ± three standard deviations contains at least 89% of the measurements in the series.

3.1.4. Shape characteristics

The moments of a rv allow to characterize its probability distribution. Moments may be computed with respect to the origin (0) or with respect to a measure of central tendency, usually the mean. The first order moment with respect to the origin is the mean of the rv and the second order moment with respect to the mean is the variance. The third and the fourth moments define the shape of the distribution.

The skewness coefficient

The skewness coefficient is the third standardized moment with respect to the mean, i.e.

$$\gamma_1 = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\sigma_x^3}, \tag{3.11}$$

where σ_x is computed as in expression (3.9). A positive coefficient means that the distribution has a long right tail, (the distribution is also known as right-skewed) while a negative coefficient means that the distribution has a long left tail (the distribution is also known as left-skewed), see for instance Figure 5. Any symmetric distribution has a skewness coefficient equal to 0, as for example the normal distribution. It should be noted though that some non-symmetric distributions could also have a null skewness coefficient.

Other statistics may also be used to detect lack of symmetry, such as the difference between the mean and the median. The mean is larger than the median in a right-skewed set of data, while it is smaller for left-skewed set of data. Positive (all values larger than 0) right-skewed sets of data do also show large standard deviations compared to their means.

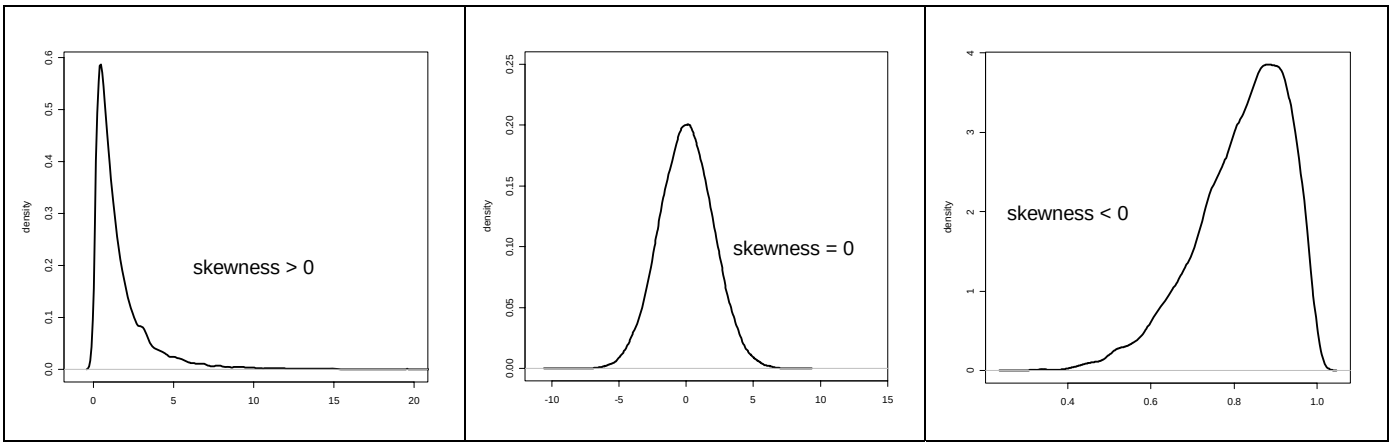


Figure 5: PDFs for distributions with different skewness coefficients: > 0 (right-skewed, left); = 0 (symmetric, center); < 0 (left-skewed, right).

The kurtosis

The kurtosis coefficient is the fourth standardized moment with respect top the mean, i.e.

$$\gamma_2 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma_x^4}. \tag{3.12}$$

It represents a measure of the "peakedness" of the distribution, see Figure 6. A kurtosis coefficient larger than 3 means that the distribution has sharper "peaks" and flatter "tails" than a normal distribution (**leptokurtic distribution**). A kurtosis equal to 3 means that the distribution is approximately normal (**mesokurtic distribution**). A kurtosis below 3 means that the distribution is flatter than the normal distribution (**platykurtic distribution**).

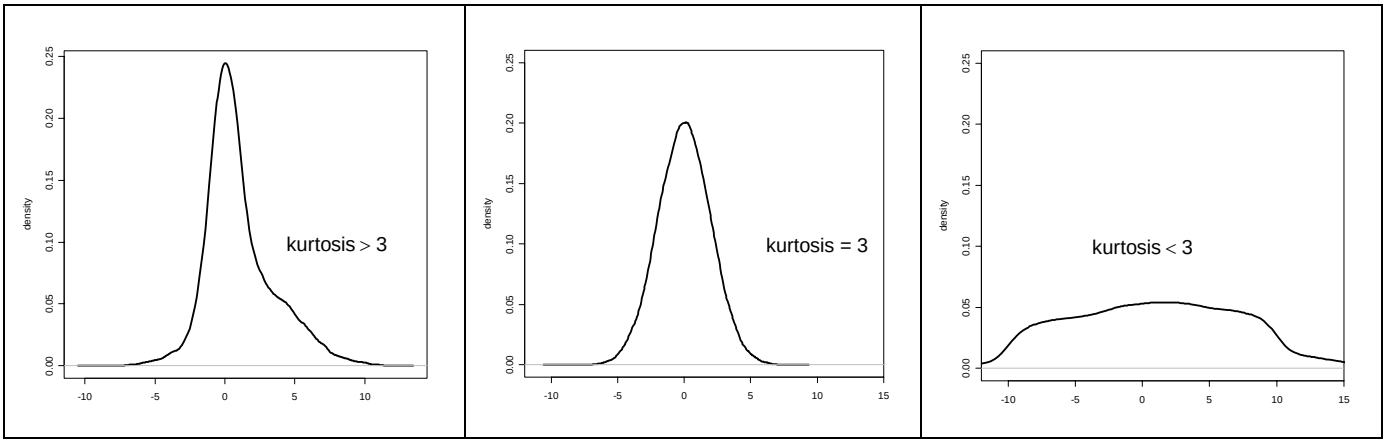


Figure 6: PDFs for distributions with different kurtosis : >3 (left) ; =3 (center); <3 (right).

The kurtosis coefficient is sometimes defined as $\mu_4/\sigma^4 - 3$, where μ_4 is the numerator in expression (3.12), in order to make the kurtosis of the normal distribution equal to zero instead of 3. The kurtosis is not an intuitive coefficient; it is quite difficult to say, by looking at a PDF if the distribution has a large or small kurtosis. What is important, in terms of shape, for a leptokurtic distribution is that there is a sharper "peak" around the mean (which means a higher probability than a normally distributed rv

of values close to the mean) and "fat tails" (which means a higher probability than a normally distributed rv of extreme value), as it can be seen in Figure 7.

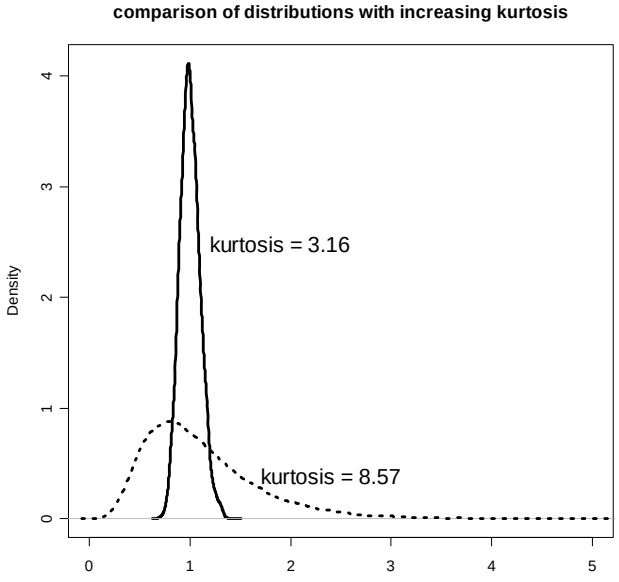


Figure 7: Influence of increasing Kurtosis

3.2. Graphical tools

3.2.1. CDF (CCDF), ECDF (ECCDF)

The cumulative distribution function (CDF) is defined by $F(x) = P(X \le x)$ and for a continuous rv it is also equal to $F(x) = \int_{-\infty}^{x} f(t)dt$, where *f*(.) is the probability density function. The most important properties of the CDF are:

- It is non-decreasing monotonic function,
- $0 \le F(x) \le 1$,
- $P(a \le X \le b) = F(b) - F(a)$,
- $F'(x) = f(x)$.

Alternatively, the complementary cumulative distribution function (CCDF), which is equal to 1-*F(x)*, may also be used. The use of the CCDF is widespread in the area of nuclear safety in general and specifically in the area of PA since many safety limits and safety criteria are given in terms of exceeding probabilities, which is the kind of information included in CCDFs.

In Figure 8 we present the CDFs and the CCDFs of some of the most frequently used distributions, without specifying their parameters, in order to give an idea of their aspects. Whenever different sets of parameters are used the position and the spread of these curves will be different.

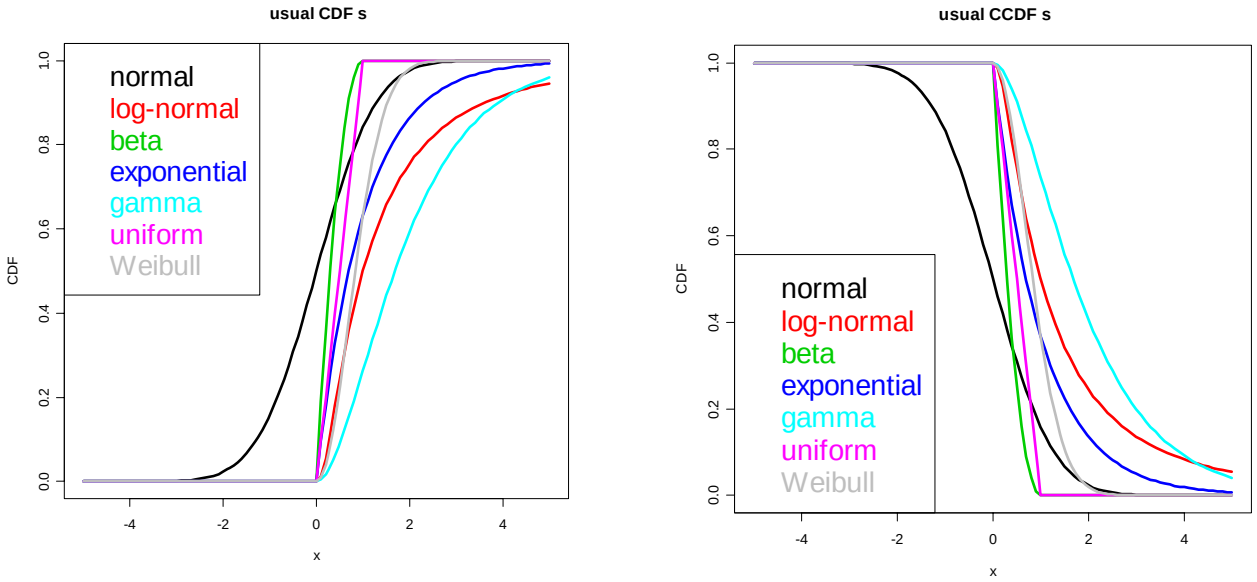


Figure 8: Some of the most usual CDFs and the corresponding CCDFs

The empirical cumulative distribution function (ECDF – $F_n(x)$) of a sample is the available tool to estimate the CDF of the corresponding rv, i.e.:

$$x \in IR,\ F_n(x) = \frac{1}{n} \#\{i, x_i \leq x\}\ , \tag{3.13}$$

where the symbol # denotes the cardinal of a set. The empirical complementary cumulative distribution function (ECCDF) is equal to $1 - F_n(x)$.

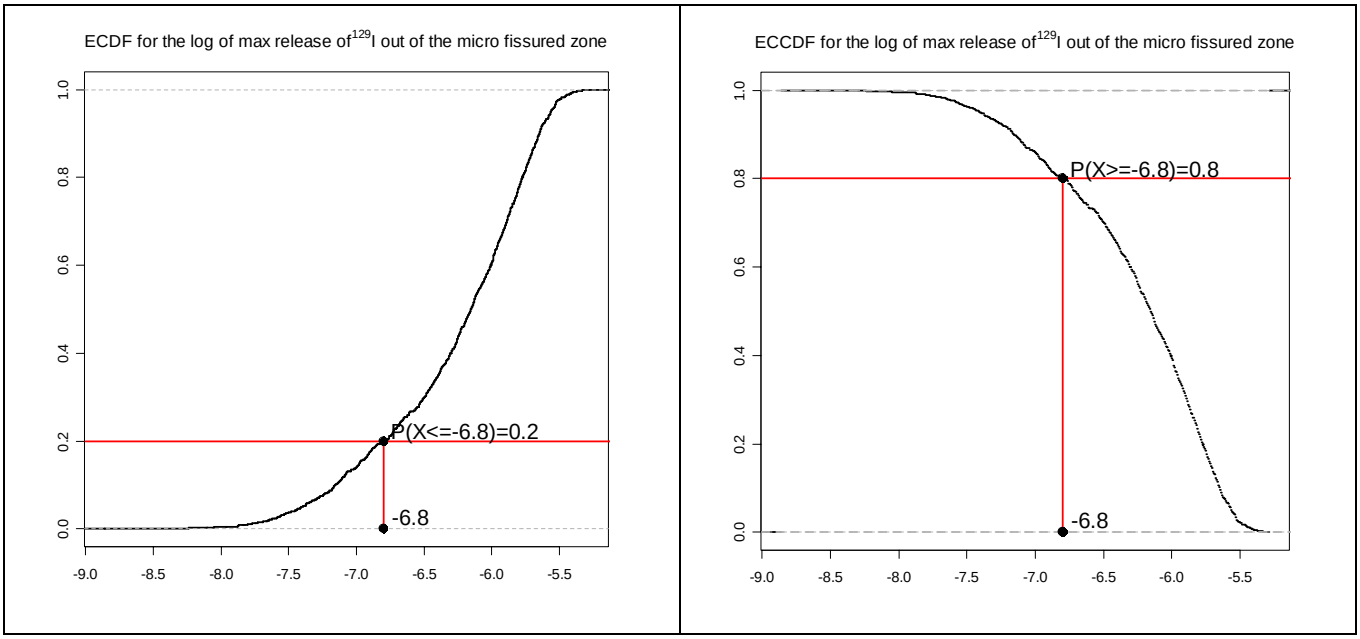


Figure 9: Example of Empirical CDF and the corresponding CCDF

In Figure 9 we represent the ECDF (left) and the ECCDF (right) for some data from the benchmark in [Prváková 08]. The sample (of size n=1000) represents the decimal logarithms of the peaks of the release of ^{129}I coming out of the disposal cell. It is easy to read directly on this representation that, for example, the percentage of the sample such that the $\log_{10}$ (^{129}I) is less than or equal to -6.8 is around

20%. The same information can be read on the right panel of Figure 9: the percentage of the sample such that the $\log_{10}$ (^{129}I) is greater than or equal to -6.8 is around 80%.

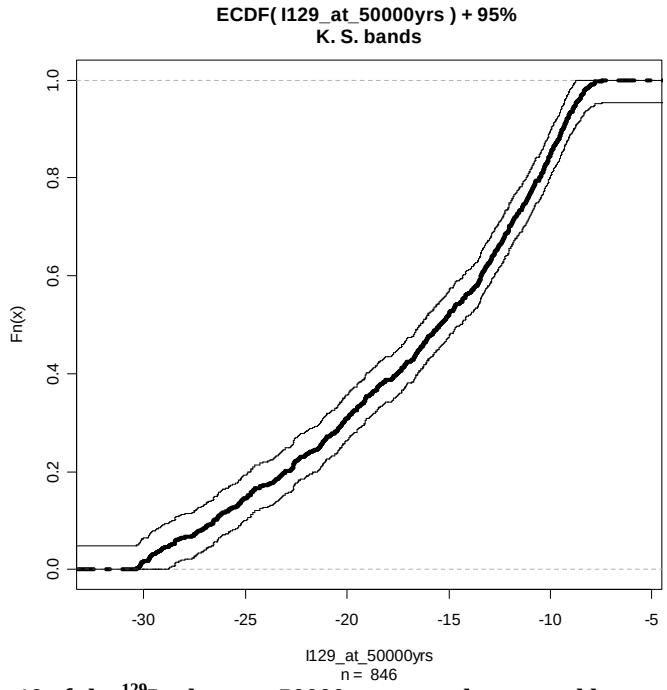


Figure 10 : ECDF for the log10 of the ^{129}I release at 50000 years, at the top and bottom of the repository, together with its 95% confidence bands.

Moreover, Kolmogorov Smirnov confidence bands may be computed for any ECDF and any ECCDF (for details on Kolmogorov Smirnov confidence bands see [Owen 01] and [Conover 80]). In Figure 10 we present an example of ECDF together with its 95% confidence bands. See section 4.1 for correctly interpreting this graphic representation.

3.2.2. Histogram

The histogram graphically summarizes and displays the distribution of a data set. The histogram is constructed by regrouping the data into *k bins* $C_1 = [a_0, a_1[, C_2 = [a_1, a_2[, ..., C_k = [a_{k-1}, a_k]$ and then defining the (relative) frequency of each bin as $f_j = \frac{1}{n} \#\{i, x_i \in C_j\}$. A density is then inferred by a step function whose value for the C_j bin is the associated frequency per unit length, i.e. $f_j / (a_j - a_{j-1})$. The surface below this step function is equal to 1. However, even if it is possible to define variable bin widths, the use of constant bin width is most popular. In the case of discrete variate two options are available: either using the cardinal of each bin (absolute frequency, see Figure 11) or using the relative frequencies. Discrete variates can also be represented as bars. Figure 12 illustrates the importance of the choice of the number of bins (or equivalently of the bins widths): the left side picture is very "noisy", too many bins have been displayed; on the contrary, the right hand side picture has not enough bins, and much of the information is therefore lost. The only reasonable histogram is the one in the middle, where the corresponding PDF (see section 3.2.3) has been added. The number of segments should be sufficient to represent the shape of the distribution but not so small so that noise becomes dominant.

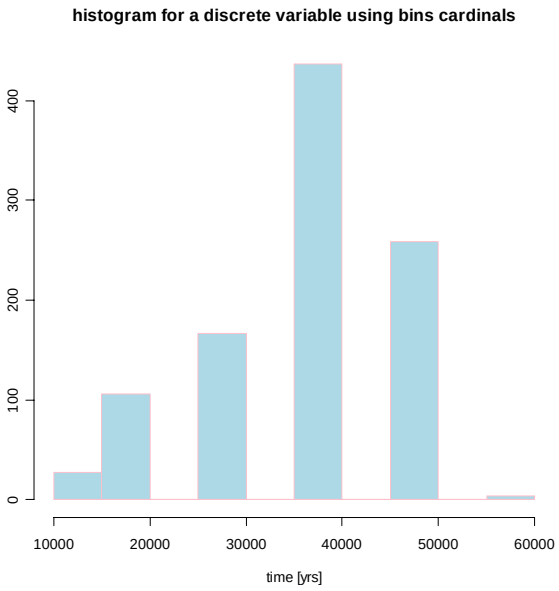


Figure 11: Histograms for a discrete variable using as ordinate the bins cardinals

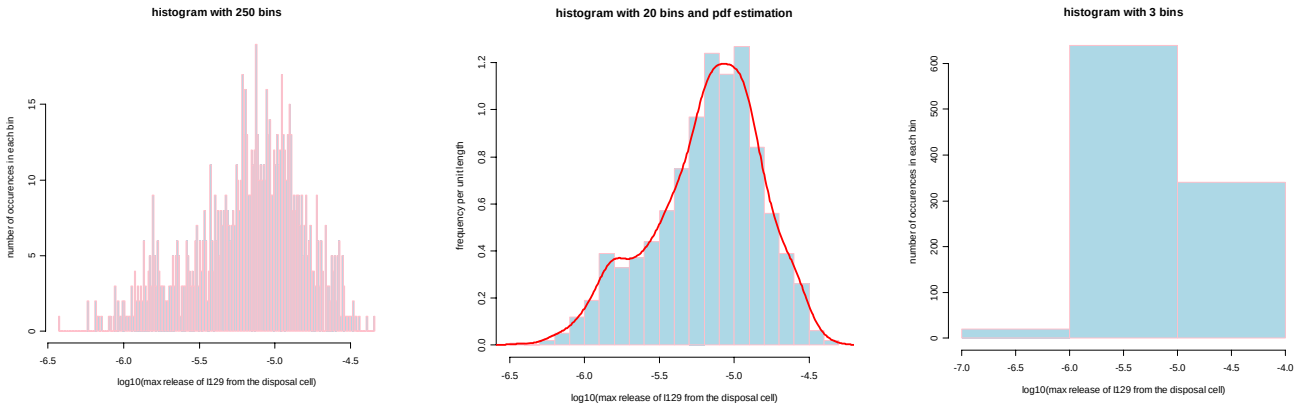


Figure 12: Histograms for the same data, using different number of bins

3.2.3. PDF estimation

Kernel method for the probability density function (PDF) estimation

The PDF represents the probability that the random variable X is in the interval $[a,b]$ in terms of integrals, i.e. it is the function f such that

$$P(a \le X \le b) = \int_a^b f(x)dx \,. \tag{3.14}$$

If we consider a sample of size n ($x_i, i = 1,...,n$) from an unknown continuous probability distribution of density f, the histogram represents an approximation of the PDF f. The main deficiencies of the histogram are its discontinuity and the appropriate choice of the number of bins (or bin widths).

The kernel estimation of the PDF is a non-parametric method (because it does not assume a certain probability distribution) generalizing the histogram. The kernel estimator of *f*, denoted by $\hat{f}$ is a sum of "bumps" of width *h* placed at the observations x_i :

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.15}$$

K denotes the kernel. There are several desirable properties:

- positivity : $K \geq 0$,
- regularity : K has to be smooth enough,
- normalization : $\int_{-\infty}^{+\infty} K(x)dx = 1$,
- symmetry : $K(x) = K(-x)$,
- fast decreasing at infinity.

The most used kernels (see Figure 13):

- Gaussian : $K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$
- Epanechnikov : $K(x) = \begin{cases} \frac{3}{4}(1-x^2), & if \quad x \in [-1,1] \\ 0 & otherwise \end{cases}$
- Rectangular : $K(x) = \begin{cases} \frac{1}{2} & if \quad x \in [-1,1] \\ 0 & otherwise \end{cases}$
- others : Triangular, Biweight, Cosine, Optcosine.

It can be seen from Figure 13 that density curves are similar for the different Kernels. Thus the kernel is not as important as the choice of bandwidth, *h*. This scaling parameter (which has the same physical dimension as the sample) controls:

- the width of the probability mass spread around a point
- the smoothness or roughness of a density estimate.

If the bandwidth is too small, the estimated density will be under-smoothed; a large value of *h,* on the other hand, would lead to an over-smoothed estimated density (see Figure 14).

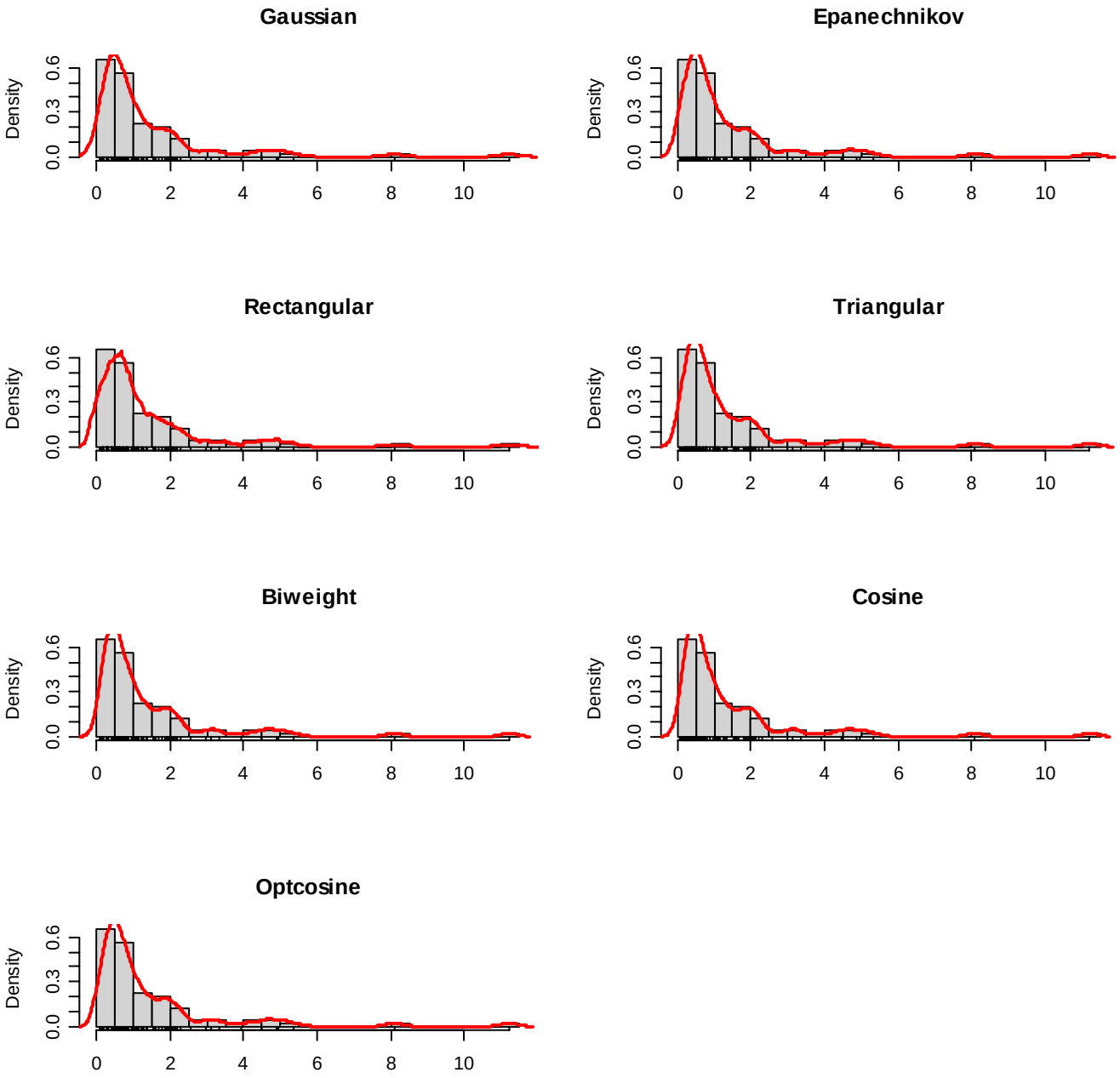


Figure 13: Influence of the kernel (h=1)

An **optimal bandwidth** may be computed for each kernel. The criterion to be minimized is either the Mean Integrated Squared Error (MISE) or the Asymptotic Mean Integrated Squared Error (AMISE). For instance, the optimal bandwidth for the Gaussian Kernel and MISE criterion is:

$$h_{opt} = 1.06\,\hat{\sigma}\,n^{-1/5} \tag{3.16}$$

where $\hat{\sigma}$ is the empirical standard deviation of the sample, i.e. $\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Unfortunately, the optimal bandwidth is over-smoothing if f is multimodal or somehow "not normal".

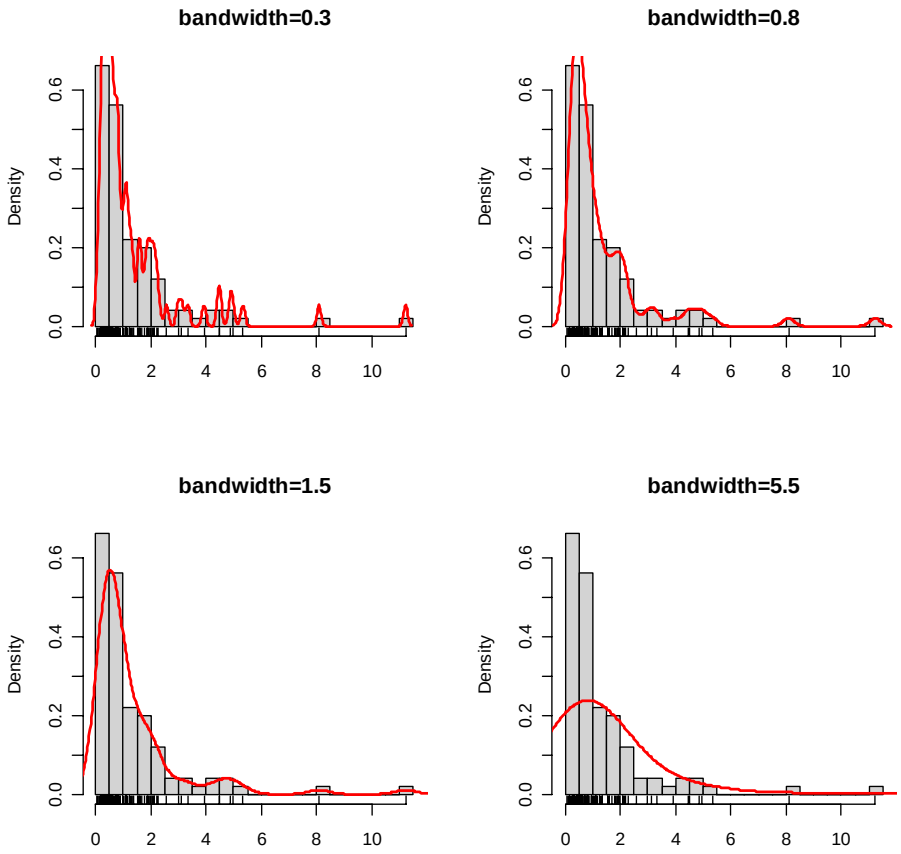


Figure 14: bandwidth influence, from under to over-smoothing (Gaussian kernel)

Another option is to use the **adaptive kernel method**, which consists of varying h with x_i, in order to have a small h where we have a high density of data and a large h where the data is sparse. The algorithm is outlined below.

1. define a pilot estimation $\tilde{f}(x)$ (an optimal bandwidth kernel estimation with optimal bandwidth denoted by h_0), such that $\tilde{f}(x_i) > 0$
2. compute $\lambda_i = \left\{\tilde{f}(x_i)/g\right\}^{-\alpha}$, where $\log(g) = (1/n)\sum \log(\tilde{f}(x_i))$ and $0 \le \alpha \le 1$ is a sensitivity parameter (a good choice is $\alpha = 1/2$)
3. $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_0\lambda_i} K\left(\frac{x - x_i}{h_0\lambda_i}\right)$.

Details concerning density estimation may be found in [Silverman 86].

3.2.4. Boxplots

A boxplot (also known as a box-and-whisker plot) is a way to picture groups of numerical data using five of their summaries (the smallest observation, lower quartile (Q_1), median, upper quartile (Q_3), and largest observation). It also indicates if there are some observations which might be considered outliers. The length of the "box" is the interquartile range (IQR = Q_3-Q_1) and the line inside the box

stands for the median. An outlier is any data that lies outside the interval $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$. The bounds of this interval are indicated by some tic marks and are connected to the box by a line.

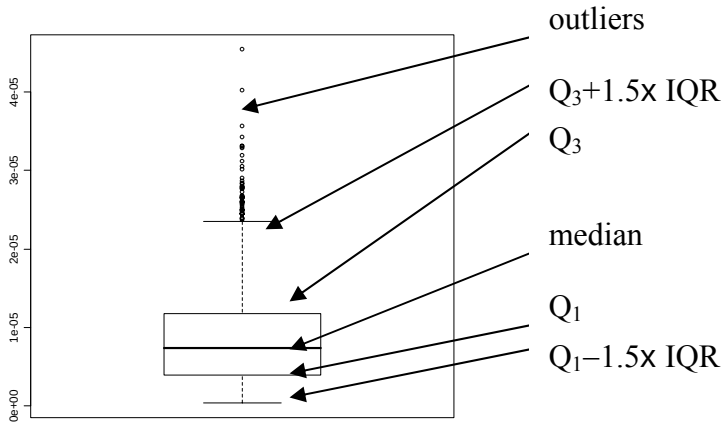


Figure 15: Example of boxplot for data representing the peak of the release of ^{129}I coming out of the disposal cell [Prváková 08].

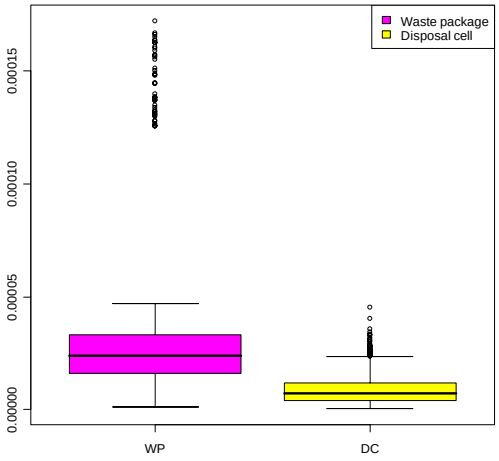


Figure 16: Comparison of the distributions for the peaks of the release of ^{129}I coming out of the waste package and of the disposal cell by using boxplots [Prváková 08].

As we can see from Figure 16, the boxplots, even if they show less information than histograms or PDFs, are very useful for making comparisons between different distributions; they may even suggest the existence of a second subpopulation instead of outliers (as it is the case for the left boxplot in Figure 16).

3.2.5. Qqplot

A qqplot is a quantile-quantile plot of two data sets and is an excellent tool for determining if the two data sets have the same parent distribution. It consists in plotting the quantiles of the first data set against the quantiles of the second data set. When two random samples have a qqplot which mostly falls on a straight line, then the two parent populations have a similar shape. It is often used to test qualitatively the conformity between an empirical and a theoretical distribution, like for instance in the linear regression context for checking the normality of the residuals.

Example:

Building a qqplot to check if a sample is normally distributed has two stages:

- order the data in ascending order : $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$
- associate to every $x_{(i)}$ the $\alpha = (i - 0.5)/n$ – the corresponding quantile (denoted here by q_α) of a centered-reduced normal distribution $N(0,1)$ and plot the couples $(q_\alpha, x_{(i)}), \forall i = 1,\ldots,n$.

If the sample is normally distributed, then the couples $(q_\alpha, x_{(i)}), \forall i = 1,\ldots,n$ should be approximately on a straight line (with slope the standard deviation of the sample).
The table below represents the quantiles of the normal distribution *N(0,1)* and some residuals from a regression on 13 observations. A quick glance at the qqplot in Figure 17 tells us that the residuals are approximately "normal", which is one of the main assumptions in a regression model.

i	q_a	$x_{(i)}$
1	-1.77	-2.16
2	-1.2	-1.51
3	-0.87	-1.11
4	-0.62	-0.76
5	-0.4	-0.74
6	-0.19	-0.62
7	0	-0.44
8	0.19	0.3
9	0.4	0.62
10	0.62	0.64
11	0.87	0.99
12	1.2	1.45
13	1.77	1.46

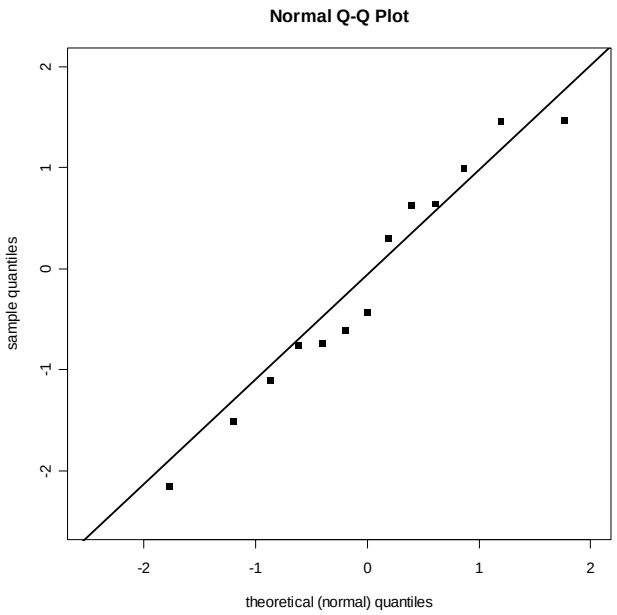


Figure 17 : The "normal" qqplot (corresponding to the data in the left table)

One should expect some variations from the straight line for small data sets.

Some other characteristic shapes for the qqplots are the following:

- when the sample distribution has a positive skewness coefficient, the normal qqplot will have a U shape as seen in Figure 18. The corresponding skewness coefficient is 1.43.

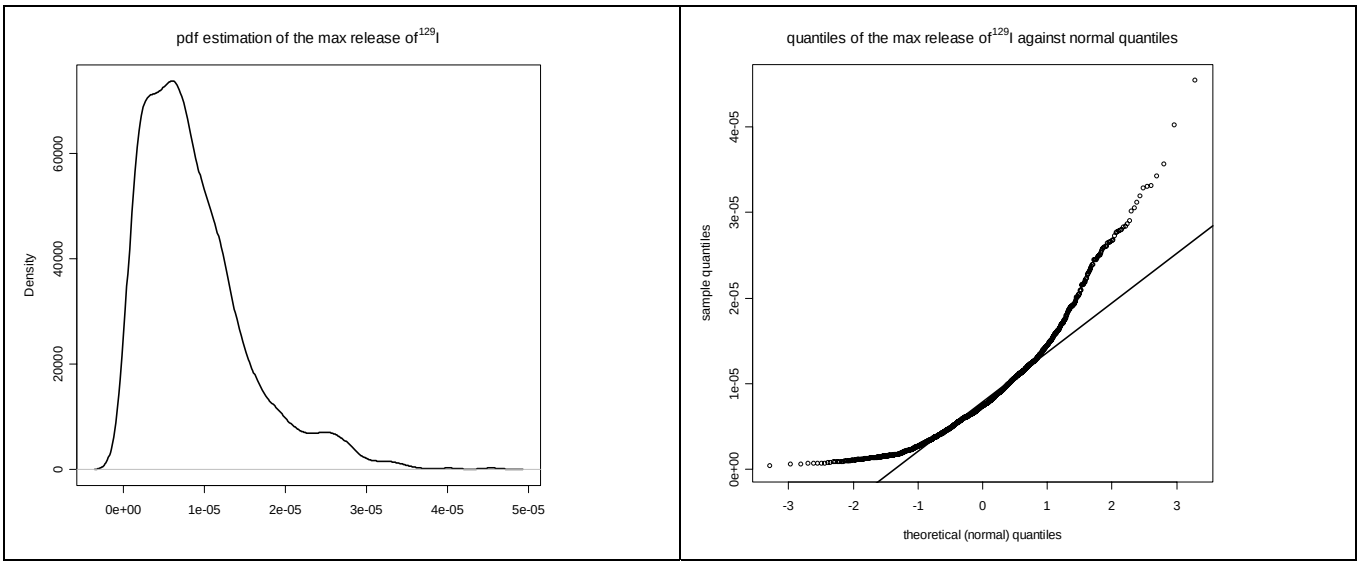


Figure 18: Example of a skewed sample distribution; on the left its PDF estimation and on the right the corresponding qqplot. The sample data is the maximum release of ^{129}I getting out of the disposal cell computed in a study of the release of radionuclides from wastes of an ILW disposal cell embedded in a porous materials for a generic French clay site [Prváková 08].

- when the sample distribution has a negative skewness coefficient, the normal qqplot will have an inverse U shape, as seen in Figure 19. The corresponding skewness coefficient is –1.43.

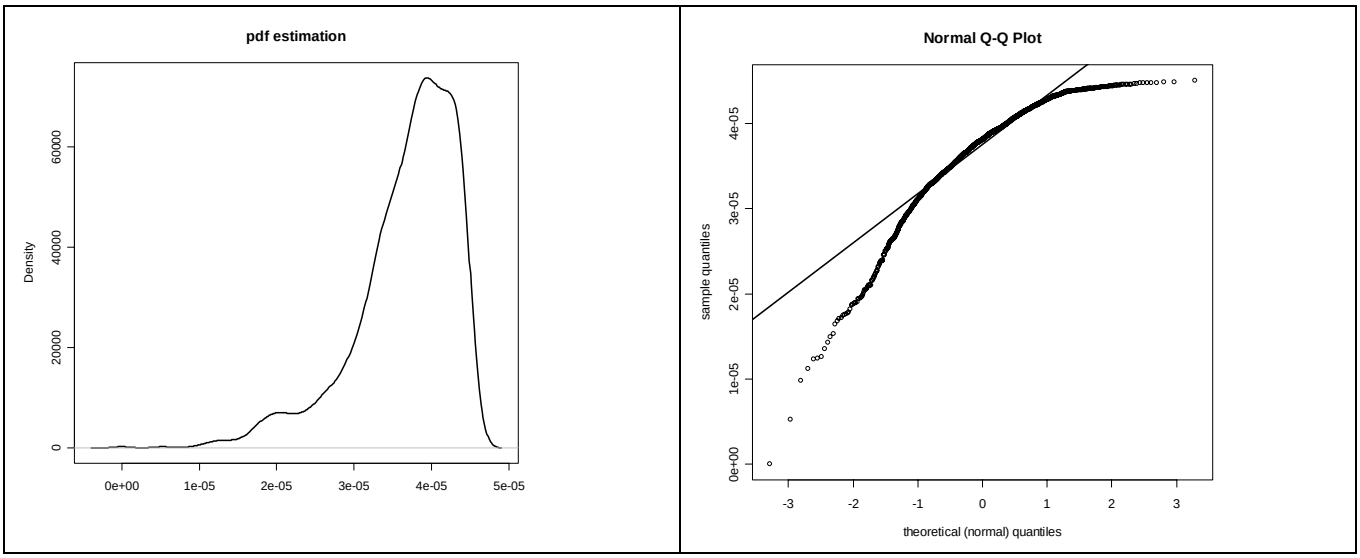


Figure 19: Example of a skewed sample distribution; on the left its PDF estimation and on the right the corresponding qqplot.

- when the sample distribution is more concentrated to the right and to the left that a normal distribution (which is a combination of the 2 previous examples), the normal qqplot will have a S shape, see Figure 20 for instance. Examples of such distributions are bi-modal and uniform distributions.

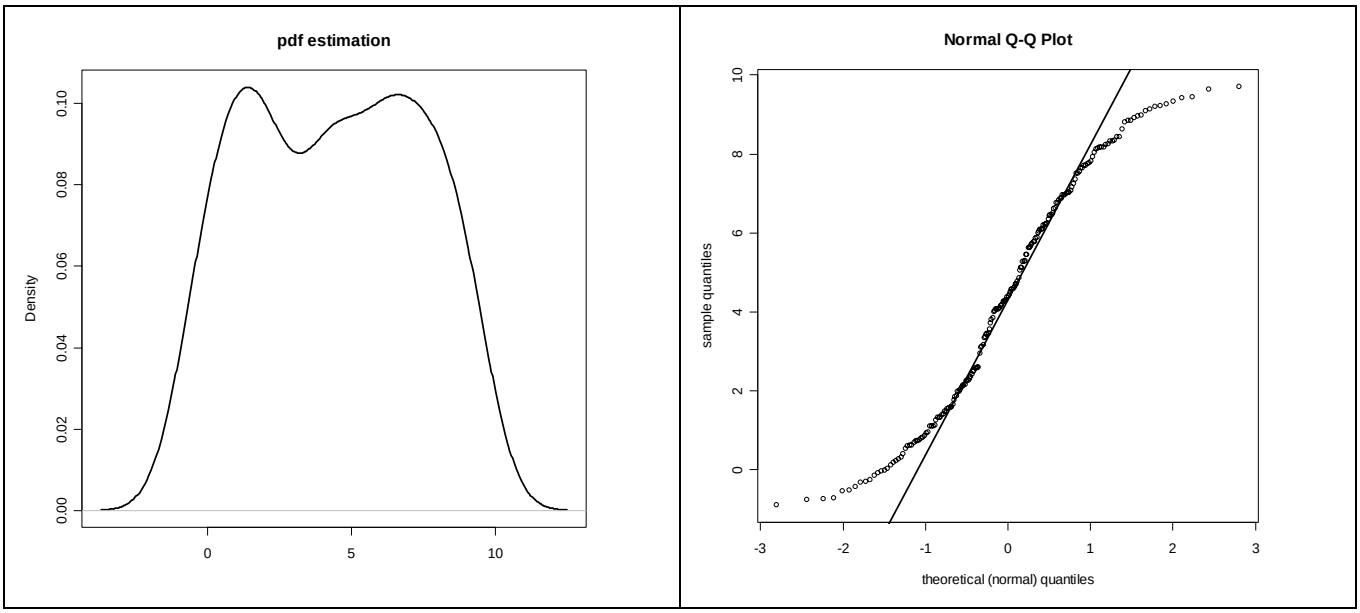


Figure 20: Example of a bi-modal distribution; to the left its PDF estimation and to the right the corresponding qqplot.

- when the sample distribution has larger tails than a normal distribution, the normal qqplot will have an inverse S shape, see Figure 21 for instance.

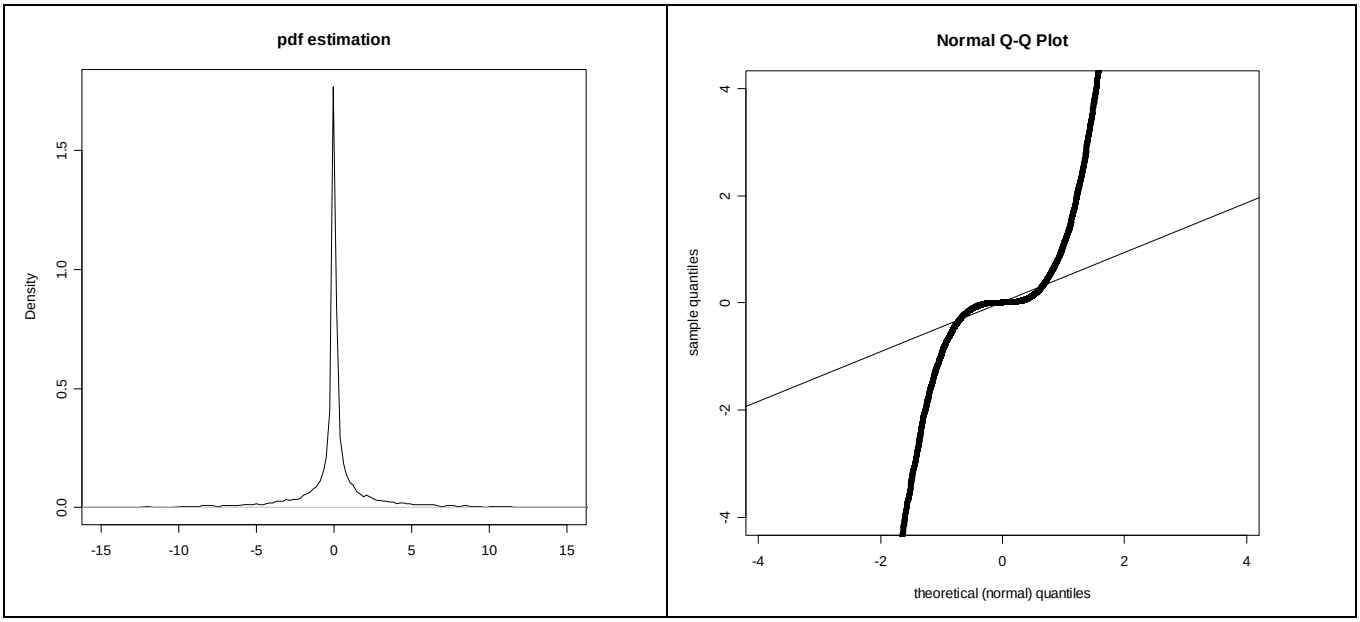


Figure 21: Example of a long tailed distribution; to the left its PDF estimation and to the right the corresponding qqplot.

- Outliers can also be detected with the qqplots: some points quite below or above the line indicate this situation. The encircled dot in Figure 22 is a very good candidate to be considered outlier.

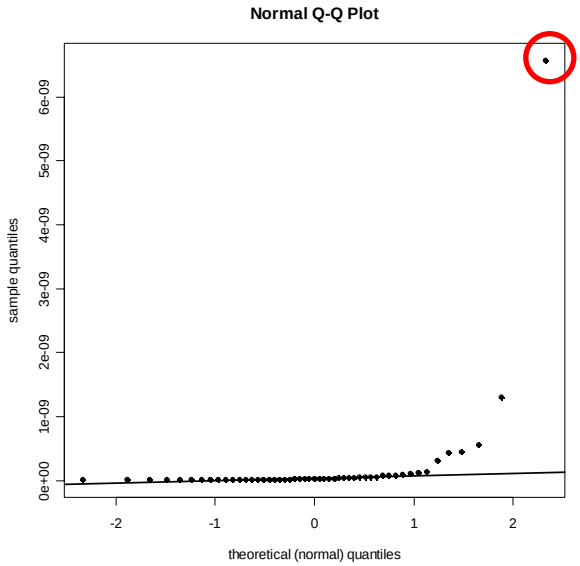


Figure 22: Qqplots for data with outliers

This approach extends straightforward for testing the conformity of samples to other distributions, and to check if two samples come from the same distribution.

4. Input uncertainty assessment

Input uncertainty assessment is the process of characterising, through probability density functions (PDFs) or probability mass functions (pmfs), the uncertainty of continuous and discrete input parameters used in PA studies. There are essentially two ways to do it: classical inference methods or Bayesian methods. Expert judgment is a third way to do it, which necessarily involves a subjectivist interpretation of probability, i.e. probability interpreted as a degree of belief about the occurrence of an event or about the truth of a proposition. The method to use depends primarily on the amount and the reliability of the data. Classical inference methods are used when a substantial quantity of data are available. Bayesian methods are used when only limited amounts of data are available. Expert judgment is used under conditions of real scarcity of data, though at least a few data should be available, otherwise any attempt of uncertainty characterisation would be pure speculation.

4.1. Classical inference methods

Classical inference methods are based on the assumption of having a random sample. The target is to determine the PDF that generated the random sample. This process may be divided in three steps:

1. Model identification
2. Parameter estimation, which is divided in two parts
    a. Point estimation
    b. Interval estimation
3. Diagnosis of the model

Model identification consists in finding the most appropriate probability model (uniform, normal, log-normal, exponential, Weibull, etc.) for the sampled data. This task needs the use of graphic tools such as histograms and qqplots, in addition to the experience in the field under study. Furthermore experts

in the field will often have an idea of the distributions that could best represent the data. This part of the process certainly involves subjective elements.

Once the probability model has been identified, the parameters need to be determined. Most probability models are characterised by a set of parameters (parametric models), as for example the mean, μ, and the standard deviation, σ, in a normal (Gaussian) probability model. Estimation is done via techniques of point estimation. These techniques allow identifying a best choice for those parameters. Identifying best choices does not mean that those are the only acceptable ones; other similar values could also be acceptable. A measure of error or of likely alternatives is also needed. This is provided by *interval estimates*.

The last step consists in checking that the hypotheses considered in the whole process were correct. Three hypotheses are normally used: the type of probability model, the independence between the different observations and the homogeneity of the sample.

In the following pages special attention will be dedicated to both types of parameter estimation (step 2) and to checking that the assumed probability model is good enough (first hypothesis tested in step 3).

4.1.1. Point estimation

There are several methods, some of them recently developed, such as Jackknife and Bootstrap methods (see [Efron 93]), but the best known and most widely used methods are the Maximum Likelihood Method and the Method of Moments. The main shortcoming of all these methods is their requirement of sample sizes to get good quality estimates. In practical situations with real engineering facilities it may be quite difficult to get the required sample size.

Method of moments

Method of Moments is probably the oldest inferential method to estimate the parameters of a PDF. K. Pearson developed the method of moments by the end of 19th century. The idea is quite simple. It consists in taking as an estimator of a parameter its equivalent sample quantity. So, the sample mean is the estimator for the mean, the sample variance is the estimator for the variance and so on.

Maximum Likelihood method

The Maximum Likelihood Method is the most widely used and most powerful estimation method in the classical context. Let us assume that we wish to study a random variable X (representing a parameter affected by uncertainty) of a known distribution function type $f(\mathbf{X}|\theta)$, but of unknown parameter θ. In order to estimate θ we take a random sample $\mathbf{X} = (X_1, X_2, ..., X_n)$, which is assumed to be a random vector, whose components are independent and identically distributed (iid), so that its joint probability density function is

$$f(\mathbf{X}|\theta) = f(X_1, ..., X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta) \,. \qquad \textbf{(4.1)}$$

It is important to notice that, in this expression, under the classical view, before sampling, θ is unknown, but has an assigned value that determines what regions of $\mathbf{X}$ are more likely and what regions are less likely. So, this is a function whose unknowns are $\mathbf{X}$. This is the meaning before sampling. As soon as the sample is available, $\mathbf{X}$ is known, while θ remains unknown. The objective is

to determine what value, among all the possible values of θ, makes the sample actually obtained the most likely one. The problem is hence to find the value of θ for which the function defined in (4.1) attains its maximum value. As it is convenient to look at the problem after getting the sample, expression (4.1) is usually written as

$$L(\theta|\mathbf{X}) = f(\theta|\mathbf{X}) = f(\theta|X_1,...,X_n) = \prod_{i=1}^{n} f(X_i|\theta)\,, \qquad \textbf{(4.2)}$$

which means that, after sampling, the probability density function of the sample vector is changed into a function of the unknown parameter θ. *'L'* stands for 'Likelihood'. From a practical point of view, the function whose maximum is actually computed is not L, but its logarithm $l(\theta|\mathbf{X})$. Both functions reach a maximum at the same point since the transformation to get one from the other one is a monotonic transformation.

As an example, let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a sample of size n of a Gaussian random variable whose variance σ^2 is known. We wish to estimate the mean μ of the random variable under study. Under these circumstances, the likelihood function is

$$L(\mu|\mathbf{X}) = f(\mu|X_1,...,X_n) = \prod_{i=1}^{n} f(\mu|X_i) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}\,, \qquad \textbf{(4.3)}$$

whose logarithm is

$$l(\mu|\mathbf{X}) = Ln(L(\mu|\mathbf{X})) = -\frac{n}{2}Ln(2\pi) - \frac{n}{2}Ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\,. \qquad \textbf{(4.4)}$$

In order to compute the value of μ for which this expression reaches a maximum, we compute its first derivative with respect to μ

$$\frac{\partial l(\mu|\mathbf{X})}{\partial \mu} = -\sum_{i=1}^{n}\left(-\frac{1}{\sigma}\right)\left(\frac{X_i-\mu}{\sigma}\right). \qquad \textbf{(4.5)}$$

Distribution	PDF	Parameters	Estimators
Uniform	$\frac{1}{b-a}; a \le x \le b$ $0; otherwise$	a: Minimum b: Maximum	$\hat{a} = \min\{x_1,...,x_n\}$ $\hat{b} = \max\{x_1,...,x_n\}$
Log-uniform	$\frac{1}{xLn(b/a)}; a \le x \le b$ $0; otherwise$	a: Minimum b: Maximum	$\hat{a} = \min\{x_1,...,x_n\}$ $\hat{b} = \max\{x_1,...,x_n\}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right];$ $\sigma > 0$	μ: mean σ^2: variance	$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$
Log-normal	$\frac{1}{x\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{Lnx-\mu}{\sigma}\right)^2\right];$ $x > 0, \sigma > 0$	μ: mean of Ln(x) σ^2: variance of Ln(x).	$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} Lnx_i$ $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Lnx_i - \bar{x}_n)^2$
Exponential	$\lambda e^{-\lambda t}; t > 0$	λ: inverse of the mean	$\hat{\lambda}^{-1} = \frac{1}{n}\sum_{i=1}^{n} t_i$
Gamma	$\frac{t^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}e^{-t/\beta};$ $\alpha > 0, \beta > 0, x > 0$	α: shape param. β: scale param.	$\hat{\alpha}\hat{\beta} = \bar{x}_n$ $\frac{1}{n}\sum_{i=1}^{n} Lnx_i = Ln\hat{\beta} + \psi(\hat{\alpha})$ *

Weibull	$\frac{cx^{c-1}}{\alpha^c}\exp(x/\alpha)^c$	α: scale param. c: shape param.	$\hat{\alpha} = \left(1/n\sum_{i=1}^{n} x_i^{\hat{c}}\right)^{1/\hat{c}}$ $\hat{c}^{-1} = \left(\sum_{i=1}^{n} x_i^{\hat{c}} Lnx_i\right)\left(\sum_{i=1}^{n} x_i^{\hat{c}}\right)^{-1} - \frac{1}{n}\sum_{i=1}^{n} Lnx_i$ †
Binomial	$\binom{n}{i} p^i (1-p)^{n-i}$	p: prob. of event	$\hat{p} = r/n$ r=number of times event happens out of n trials
Poisson	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ: Mean n. of events per unit of time, length, surface, etc.	$\hat{\lambda} = r/n$ r= number of events n= sample size (s, m, m^2, etc.)

Table 1: The most useful probability distributions functions, their parameters and their maximum likelihood estimators.
* The solutions of this system of equations, where ψ stands for the digamma function, are the maximum likelihood estimators.
† c is estimated recursively from the second equation, later on its estimate is substituted in the first one in order to get the estimator of α.

The maximum is obtained when this expression equals zero, which happens for the value

$$\hat{\mu} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ .} \qquad \textbf{(4.6)}$$

The reader may check, by computing the second derivative that, indeed, the likelihood function reaches a maximum when $\mu = \hat{\mu}$ (second derivative is less than zero when $\mu = \hat{\mu}$). The method may also be applied when a PDF is defined through a vector of parameters; in that case the usual rules for maximizing a multi-parameter function must be applied (to equal first partial derivatives to zero and to check conditions imposed on the Hessian matrix evaluated at the point where first partial derivatives are zero). The method provides a single value as an estimate. If needed, a confidence interval with the desired degree of confidence, may be obtained using interval estimation theory.

The maximum likelihood method has several properties that makes it the most widely used estimation method [Mood 74]:

- The estimators obtained through this method are asymptotically unbiased (the limits of their expected values when the sample size tends to infinite are the true values of the parameters).
- They are asymptotically normal since their distributions become normal when the sample size tends to infinite.
- They are asymptotically efficient; for large sample sizes, they are the most accurate estimators.
- They are sufficient since they summarise all the relevant information contained in the sample.
- They are invariant; if $\hat{\theta}$ is the maximum likelihood estimator of θ, and $\theta' = f(\theta)$, then $f(\hat{\theta})$ is the maximum likelihood estimator of θ'.

4.1.2. Interval estimation

The purpose of point estimation is to give some single "best" value of each unknown parameter, based on sample data. Nevertheless, any point estimate cannot completely describe the distibution. Due to the way the estimation process is conducted, the estimate and the actual value of the parameters are close, but they are usually different. Scientists and engineers try to provide, at least, a measure of the error made when a point estimate is given. Interval estimation was created to solve this problem.

Confidence intervals are the main tool to estimate intervals for a given parameter in a probability model. The theory of confidence intervals is based on the study of the distribution of the sample mean, the sample variance and other statistics and on the concept of pivotal quantity. If we take a sample of size n from a Gaussian variable and we compute the sample mean we will get a given value, usually

close to the mean μ of that variable. If we get a new sample of size n, we can compute a new sample mean. We may repeat the same process k times and we will get a sample of size k of the sample mean based on n observations. By plotting these k values as a histogram, we will get an idea of the distribution with the associated sample mean. Any standard statistics book (see [Mood 74] or [Casella 90]) shows that, for a Gaussian variable, the sample mean follows a Gaussian distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. The sample mean as a random variable has the same mean as the variable itself but its standard deviation is smaller. In fact, the larger n, the smaller its standard deviation. Additionally, its distribution is also normal. Taking into account the properties of normal distributions, this means that the quantity $(\overline{X}-\mu)/(\sigma/\sqrt{n})$ follows a standard Gaussian distribution with mean=0, standard deviation=1. This quantity is referred to as 'pivotal quantity'; it is a function of the sample values and the parameter studied but whose distribution does not depend on the actual value of the parameter. Knowing the distribution of this pivotal quantity, we obtain

$$P\left[-z_{\alpha/2} \le (\overline{X}-\mu)/(\sigma/\sqrt{n}) \le z_{\alpha/2}\right] = 1-\alpha \Leftrightarrow P\left[\mu \in \overline{X} \pm z_{\alpha/2}(\sigma/\sqrt{n})\right] = 1-\alpha \quad , \tag{4.7}$$

where $z_{\alpha/2}$ stands for the 100(1-α/2)% percentile of the standard Gaussian distribution. Expression (4.7) means that the interval

$$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{4.8}$$

is a 100(1-α)% confidence interval for the mean of that normal distribution whose standard deviation is known. Typically α is set to 0.05 and then $z_{\alpha/2}$=1.96. In this case the interval obtained is a 95% confidence interval.

Distribution	Parameter	Pivotal quantity	Distribution of the pivotal quantity	Confidence interval
Normal	μ (σ^2: known)	$(\overline{X}-\mu)/(\sigma/\sqrt{n})$	Standard Gaussian: N(0,1)	$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Normal	μ (σ^2: unknown)	$(\overline{X}-\mu)/(\hat{S}/\sqrt{n})$	t_{n-1}	$\left[\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Normal	σ^2	$(n-1)\hat{S}^2/\sigma^2$	χ^2_{n-1}	$\left[(n-1)\hat{S}^2/\chi^2_{\alpha/2}, (n-1)\hat{S}^2/\chi^2_{1-\alpha/2}\right]$†
Exponential	λ	$2n\lambda\overline{X}$	χ^2_{2n}	$\left[\chi^2_{1-\alpha/2}/(2n\overline{x}), \chi^2_{\alpha/2}/(2n\overline{x})\right]$†
Generic	θ	$\frac{\theta-\hat{\theta}_{ML}}{\sigma(\hat{\theta}_{ML})}$	*Standard Gaussian: N(0,1)	$\left[\hat{\theta}_{MV} - z_{\alpha/2}\sigma(\hat{\theta}_{MV}), \hat{\theta}_{MV} + z_{\alpha/2}\sigma(\hat{\theta}_{MV})\right]$

Table 2: Confidence intervals for normal, exponential and generic probability distributions**.**
† $\chi^2_{\alpha/2}$ stands for the 100(1-α/2)% percentile of the corresponding χ^2 distribution (i.e. with as many degrees of freedom as indicated in the fourth column of the table)
* Stands for asymptotic results, which means that they are valid for large sample sizes; all the others are exact results.

Interpretation of confidence intervals

Suppose that a pivotal quantity is used to estimate a 100(1-α)% confidence interval $[\theta_1, \theta_2]$ for a given parameter θ of a probability model according to the procedure above described. A priori, the probability that the interval $[\theta_1, \theta_2]$ contains θ is 100(1-α)%. The values θ_1 and θ_2 are computed on a sample; once they are computed, the true value of the unknown parameter is either in the interval

$[\theta_1, \theta_2]$ or outside it, hence we cannot speak about probability any more. By repeating the experiment (i.e. by taking different samples and by computing the interval $[\theta_1, \theta_2]$) a certain number of times, in average 100(1-α)% of the cases, the true parameter will be in the confidence interval. But we don't know in which cases this will happen. This is the reason why the well-known expression "with confidence 100(1-α)% the parameter lies in the confidence interval" is used. Figure 23 shows the results of generating via sampling 48 95% confidence interval. Only three of them do not contain the real value of the parameter (dashed line), which is close to what would be expected, between 2 and 3 intervals should not contain the real value (5% of 48).

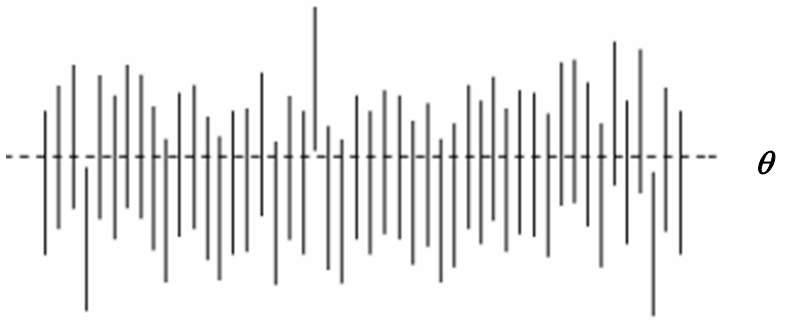


Figure 23: Repeated confidence interval (vertical lines) together with the true value of the parameter (horizontal line)

The main problem related to the use of confidence intervals is that exact confidence intervals are available only for the parameters of a few distributions such as normal, log-normal and exponential distributions. For any other distribution, only approximate confidence intervals are available, which are based on the asymptotic normality and lack of bias of maximum likelihood estimators. Table 2 shows the most frequently used confidence intervals. Exact interval estimates are available for quantiles of any distribution, provided that large enough samples are available (see section 5.4.3).

4.1.3. Goodness of fit tests

The last step of the inferential process is to check if the hypotheses under which it has been developed are true. The main hypothesis is the selected probability model. After selecting the model, the point estimation gives the best choices for the values of the parameters subject to some criteria (maximisation of the likelihood function or some other one). Both sets of information define completely the law that supposedly generated the data under study. Nevertheless, the best choice could be 'not good enough'. This is what we try to find out using goodness of fit tests. The main tests are the χ^2 (chi-square) test and Kolmogorov's test.

χ^2 (chi-square) test

The χ^2 test is based on the comparison of the histogram of the data with the estimated PDF. It consists of the following steps:

1. group the data in k sets as done when drawing a histogram and count the number of data in each set (O_i),
2. compute the probability of each set (p_i) under the assumed probability law. Compute the expected number of data in each set under the assumed probability distribution using the formula $E_i = np_i$,
3. compute the discrepancy between what is expected under the assumed model and what has been obtained in the sample according to

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

4. compare this value with the 1-α quantile of the χ^2_{k-r-1} distribution (χ^2_α). Typically α is set to 0.05 or 0.01.
    a. if $\chi^2 > \chi^2_\alpha$, reject the null hypothesis, which means that the PDF obtained though the estimation process and the data differ so much that it is very unlikely (probability < α) that the data could have been generated under the estimated distribution.
    b. if $\chi^2 \leq \chi^2_\alpha$, accept the null hypothesis. In this case the agreement between the estimated PDF and the data is good enough to consider that the PDF could have likely generated the data.

Here $k - r - 1$ is the number of degrees of freedom of the χ^2 distribution taken as a reference in the test; r is the number of parameters of the PDF that were estimated from the data to determine the PDF. So, if we consider that a given set of data could follow a normal distribution whose mean is unknown but whose variance is known. To define the PDF completely we estimate only the mean from the data. In this case r=1. If we estimate both the mean and the variance from the sample, r would be 2. The χ^2 test is an asymptotic test, it works well with large sample sizes, but it is not recommended to apply it to small data sets (in fact many authors discourage its use when the sample size is below 25 or 30).

Kolmogorov's test

Kolmogorov's test is based on the comparison of the ECDF obtained from the data and the estimated CDF. The steps to perform the test are:

1. draw the ECDF based on the data,
2. draw the CDF according to the model selected and the estimated parameters,
3. compute the maximum vertical distance (D_n) between the ECDF and the CDF,
4. compare this value with the 1-α quantile ($D(n)_\alpha$) of Kolmogorov's statistic (D(n)) distribution for a sample of size n. As usual, α is set to 0.05 or 0.01.
    a. if $D_n > D(n)_\alpha$, reject the null hypothesis, which means that the CDF obtained though the estimation process and the data differ so much as to consider very unlikely (probability < α) that the data could have been generated under the estimated distribution.
    b. if $D_n \leq D(n)_\alpha$, accept the null hypothesis. In this case the agreement between the CDF and the data is good enough as to consider that the CDF could have likely generated the data.

Kolmogorov's test is an exact test that can be applied to any random sample, whatever its size is, though its capability to detect departures from the null hypothesis is quite limited for small sample sizes.
An alternative test when the population parameters are not specified is Lilliefors test (see [Lilliefors 67]).

Example:

A random sample of size 100 has been obtained. We assume that it comes from a uniform distribution defined in the interval [0,1]. In order to test this hypothesis we perform the χ^2 test and the Kolmogorov test. In order to perform the former, we plot a histogram of the data set, which is shown in Figure 24, and compare it with what would be expectable from the theoretical PDF (see horizontal line at height 10). Then we compute the quantity $\chi^2 = [(11-10)^2/10+(5-10)^2/10+(13-10)^2/10+\ldots+(10-10)^2/10]=5.4$. Since $\chi^2 = 5.4 \leq 16.9 = \chi^2_{0.05}$ (the value of the statistic chi-square does not exceed the 95% percentile of the χ^2_9 - chi-square distribution with 9 degrees of freedom), the null hypothesis (the data set comes from a uniform distribution defined in the range [0,1]) is accepted.

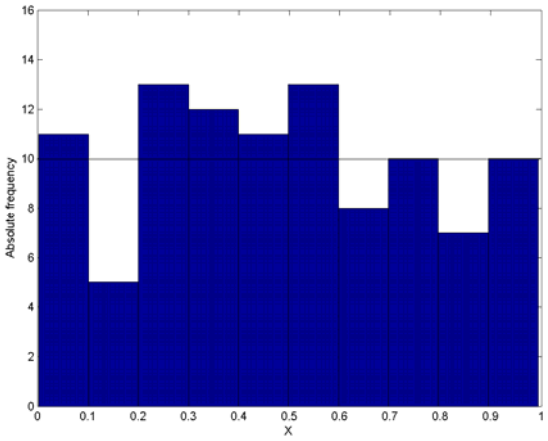


Figure 24 : Histogram obtained from random sample of size 100 and the corresponding theoretical absolute frequency (for uniform distribution).

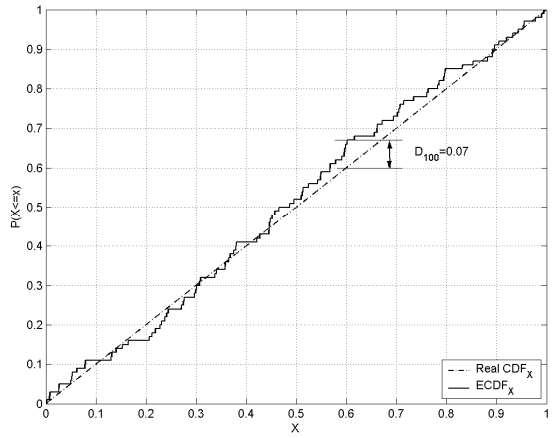


Figure 25 : ECDF obtained from a random sample of size 100 and the corresponding theoretical CDF.

In order to apply Kolmogorov's test to the same data set, we draw the ECDF and the CDF, and we compute the maximum vertical distance between both curves, see Figure 25. We then compare the value D_{100} obtained with the 95% percentile of Kolmogorov's statistic for sample size 100. Since $D_{100}= 0.070 \leq 0.0136 = D(100)_{0.05}$, we accept the null hypothesis.

4.2. Bayesian inference methods

The Bayesian interpretation of probability makes Bayes' formula a powerful tool to update degrees of belief when new information is available about an event or a proposition. Let H be the knowledge of a person (expert), and let $\{z_i\}_{i\in I}$ be a partition of the sample space of events. The Bayesian probability attributed by an expert to a given event z_k is $P(z_k|H)$. The acquisition of a set of new evidence H' induces a change in the probability given by Bayes' formula

$$P(z_k|H,H') = \frac{P(H'|H,z_k)\cdot P(z_k|H)}{P(H'|H)}, \tag{4.9}$$

where $P(z_k|H,H')$ is the 'a posteriori' probability of z_k, $P(z_k|H)$ is the 'a priori' probability of z_k and $P(H'|H,z_k)$ is the likelihood of evidence conditional on the knowledge H and the occurrence of event z_k. $P(H'|H)$ is the probability of new evidence conditional on previous knowledge, which may be considered a normalising factor, since the sum of expressions like (4.9) over the whole partition must be 1 (equivalently, the sum of the a posteriori probabilities of all the partition elements must be 1). That probability is given by

$$P(H'/H) = \sum_i P(H'/H,z_i)\cdot P(z_i/H), \tag{4.10}$$

and may be ignored in any intermediate computation. So, equation (4.10) may be written as

$$P(z_k/H,H') \propto P(H'/H,z_k)\cdot P(z_k/H), \tag{4.11}$$

which means that the a posteriori probability is proportional to the a priori probability and to the likelihood of evidence.

Two remarkable observations can be inferred from (4.9). If the a priori probability of an event is zero, the a posteriori probability will remain zero, even though the evidence against it could be very strong. So, much care should be taken when providing a priori probabilities. Null a priori probabilities should be avoided, unless total evidence about the impossibility of the events or propositions under study is available. In English literature this is called Cromwell's statement. The second result is related to the existence of strong evidence. In that case, likelihood will be completely dominant and the a priori probability will be almost irrelevant (a posteriori probability and likelihood will be almost equal). This is the case of large sample sizes, for which relative frequencies and Bayesian probabilities will be almost equal.

Bayesian inferential methods are most used under conditions of scarcity of data. The main steps of the formal process are similar to the steps of a classical inferential process: The selection of the probability model, the estimation of parameters and the diagnosis of the model. The main difference is in the estimation process, which is subject to the use of Bayes formula, as explained above. The next paragraphs gives an example of Bayesian estimation.

Let us assume a random variable X whose PDF is $f(X|\theta)$. This PDF is completely defined by the parameter θ, which is unknown but we want to estimate it. In order to start this estimation process, under the Bayesian framework, the parameter θ is considered as a random variable characterised through an a priori distribution $\pi(\theta|H)$. The a priori distribution provides information about the values the person/expert expects θ would likely take. In order to improve our knowledge about θ, we take a sample - evidence - $\mathbf{X}=(X_1,X_2,...,X_n)$, which will have $P(\mathbf{X}|\theta,H)=\prod_{i=1}^{n} f(X_i|\theta)$ as a likelihood function. Applying Bayes' formula provides the a posteriori distribution to be assigned to θ:

$$\pi(\theta|\mathbf{X},H) \propto P(\mathbf{X}|\theta,H)\cdot\pi(\theta|H), \tag{4.12}$$

which is a new PDF.

Let us assume the specific case of a Gaussian random variable X. Let us also assume that we do not know its mean, μ, though we know its variance, σ^2. Let us further assume that, given our knowledge about it, we think that μ should have a value close to μ_0, let us also assume that μ could be, equally likely, larger or smaller than μ_0, and the further away from it the less likely. Under these conditions, a Gaussian a priori PDF for μ, with mean μ_0 and variance σ_0^2, could be justified. So that $\pi(\mu|H)\sim N(\mu_0,\sigma_0^2)$. Given a sample taken from the studied variable, its associated likelihood would be:

$$P(\mathbf{X}|\mu,H)=(2\pi\sigma^2)^{-n/2}\cdot e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}. \tag{4.13}$$

When putting this expression into (4.10) and after some computations, we obtain as an a posteriori distribution for μ

$$\pi(\mu|\mathbf{X},H)\sim N(\mu_n,\sigma_n^2), \tag{4.14}$$

where μ_n and σ_n^2 are

$$\mu_n = \frac{\frac{n\overline{X_n}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \qquad \text{and} \qquad \sigma_n^{-2} = n\sigma^{-2} + \sigma_0^{-2}. \tag{4.15}$$

A priori, μ was considered to take values around μ_0, while after getting the information contained in the sample, values considered likely are those around μ_n. Additionally, the larger the sample size *n*, the closer μ_n and the sample mean, $\overline{X_n}$, will be (the larger n the larger the information contained in the sample is, while the a priori information remains constant). σ_n^{-2} is the accuracy of the estimation (the sum of the accuracy of the a priori distribution, $1/\sigma_0^2$, and the sample accuracy, n/σ^2). The larger the a priori knowledge and the larger the sample size are, the larger the accuracy (the smaller the variance) of the a posteriori knowledge about μ is. Figure 26 shows the normalised likelihood, and the a priori and the a posteriori PDFs assuming the following data: $\sigma^2 = 2$, $\mu_0 = 14$, $\sigma_0 = 2$ and $\mathbf{X} = (-3, 15, 23, 8, 13, 17)$. As previously described, the mean of the a posteriori distribution, $\mu_n = 12.43$, is between the mean of the a priori distribution, $\mu_0 = 14$, and the point where the likelihood function reaches its maximum, $\overline{X_n} = 12.17$. With the classical Maximum likelihood Method described in section 4.1, the estimate would be the one provided by (4.6): $\hat{\mu} = \overline{X_n} = 12.17$.

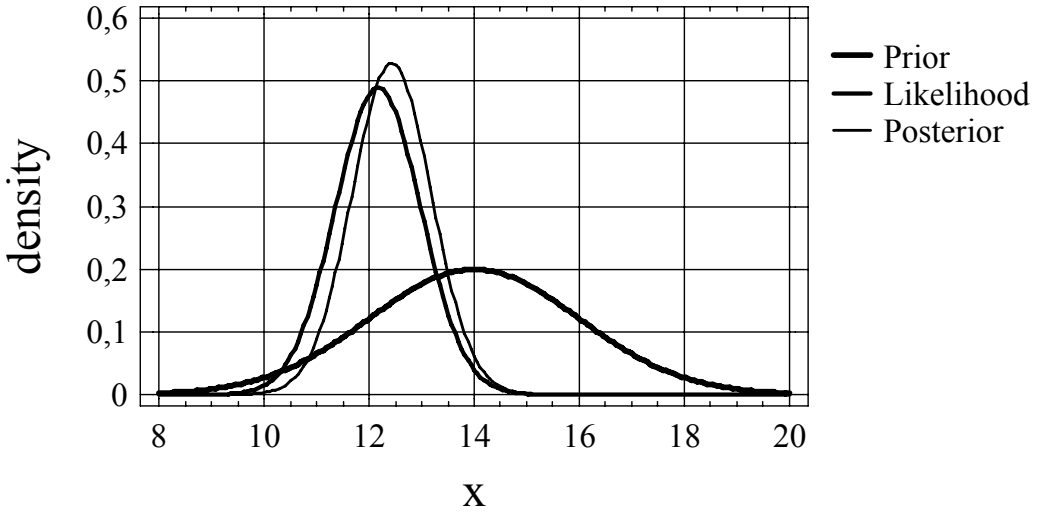


Figure 26 : A priori distribution, a posteriori distribution and likelihood function for the example of estimation of a Normal (Gaussian) random variable.

The validity of this estimation method is supported by:

1) its consistency with the way human beings learn from experience,
2) by its convergence to the results provided by the Maximum Likelihood Method when the sample size increases (after analysing the first expression in (4.15), the reader may check that when sample size increases, the mean of the a posteriori distribution converges to the sample mean, that is the estimator of μ provided by the Maximum Likelihood Method), independent of the election of the a priori PDF, except in the aforementioned case of null a priori probabilities.

One important issue that deserves being remarked is the natural and simple way of producing interval estimates in the Bayesian framework. Since the parameters used in probability models are random

variables themselves, an interval may be immediately generated as soon as the a posteriori PDF is generated,. In the example provided above, the a posteriori PDF for the mean was a Gaussian distribution with mean $\mu_n = 12.43$ and standard deviation $\sigma_n = 0.56$. This means that μ takes values in the interval $\mu_n = 12.43 \pm 1.96 \times 0.56$ with probability 0.95, or equivalently, the interval [11.33,13.53] is a 95% interval for μ. It is important to pay attention to the fact that the expression used is 95% interval, not 95% confidence interval. It is a 95% interval because, according to the probability model followed by μ, it takes values in this interval with that probability. No concept like the 'pivotal quantity' or the sampling distribution of any statistic has been used to generate such interval. The 95% confidence interval obtained for μ using expression (4.8) would be [11.03,13.31], whose interpretation was given at the end of section 4.1.2.

4.3. Expert judgment

The use of Expert Judgment (EJ) techniques is unavoidable in the performance assessment of a HLW repository due to the lack of data of many of the involved phenomena. In some cases, it is almost impossible, from a physical point of view, to get the data we need to feed our computer codes, in other cases the cost of getting them is so high that only a very small data set may be obtained. In what follows, we will list the steps of a generic EJ protocol based on most widely known protocols available in scientific literature (see for example [NUREG-1150]):

1. Selection of team project,
2. Definition of the questions to be studied,
3. Selection of experts,
4. Training,
5. Tasks definition,
6. Individual experts' work,
7. Elicitation of experts' opinions,
8. Analysis and aggregation of results,
9. Documentation.

Comprehensive information may be obtained in references such as [Mengolini 05] and [Simola 05].

Project team selection

The project team consists of analysts and generalists. Analysts are in charge of organising the steps of the protocol, so, they should have a sound background in Probability and Statistics Theory, in Knowledge Psychology and in elicitation techniques. Additionally, they should be skilful at working with people, since they will have to extensively interact with experts. The number of analysts needed depends on the extent and scope of the EJ application, though usually a couple of analysts is enough, even for large applications. Generalists provide help to the analysts in all subjects related to the specific area of knowledge of the problem to be solved. They should be able to help experts when decomposing a problem and they should be skilful at getting information sources as needed. So, they should have a good general knowledge about the problem at hand, though they do not need to be leading experts in that field. The organisation interested in the study usually provides the generalists.

Definition of the questions to be studied

Once the project team has been made, analysts and generalists must define the questions to be evaluated by the experts. The starting point for any question to be solved is usually vague. It is necessary to arrive at a *complete definition* of the uncertainty parameters we want to characterise. *Complete definition* of a parameter means the full definition of the parameter, the initial conditions for

evaluating it and any other implicit hypothesis under the initial conditions. The final definition must be unambiguous.

The complete definition of the question includes the way the experts should provide their answers. Since most of the uncertainties that come up in a PA study are characterised as Bayesian Probabilities, experts should provide their assessments of uncertainty through this kind of probabilities. So experts should provide probability distributions, either discrete or continuous.

After the full definition of the question, a list with all relevant sources of information should be done. Potential decompositions of the parameters could be done. The list of references to be considered in the list must show the actual state of knowledge in that area, but independence and reliability of the sources should always be kept in mind. When experts are expected to use computer codes for their assessment, the project team should foresee the potential training of experts in uncertainty propagation techniques (sampling, response surfaces, estimation, order statistics, etc.).

Selection of experts

The only objective of this phase is to select the most qualified experts to perform the assessment. Qualified Experts are those that:

1. Have the necessary knowledge and experience to perform the assessment,
2. are willing to participate in the assessment, and
3. do not have important motivational biases.

The first step to get the final list of experts is to start with a large list of potential experts. That first list could be based on the opinion of the generalist plus a thorough search in the scientific literature in that area. A screening should be done checking the three points in the list above. If necessary, interviews should be done to check those conditions, mainly the third one. After performing the screening, a shorter list should be obtained, from which the final selection of experts will be done. In order to arrive at the final list, two criteria should be taken into account: The number of experts to assess each question should preferably be between three and five (based on Bayesian combination of opinions' criteria) and the experts should have as much diversity as possible (different background, working at different types of organisations, etc.).

Training

The objective of this phase is to let experts know normative aspects of EJ elicitation processes. This main objective may be decomposed as the following sub-objectives:

1. motivate experts to provide rigorous assessments,
2. ensure they have knowledge of basic concepts of Probability and Statistics,
3. provide them training in the assessment of Bayesian probabilities, and
4. ensure their awareness of basic issues related to knowledge biases.

During the motivation phase the experts must get information to point out the importance of the work they are going to do. Firstly, the project team explains to the experts the framework where their opinions will be used, stressing the part of the study where their opinions are relevant. Secondly, the necessity of EJ will be explained, letting them be aware of the concept of *Lack of Knowledge Uncertainty*, and how it links to them. Thirdly, the project team will let them know that the key issue is not to predict a single value of each parameter under study, but characterising their uncertainty, allowing others to know the actual state of knowledge in that area.

After remembering basic Probability and Statistics concepts, the experts get some training on assessing Bayesian probabilities, which includes: Accurate definition of questions to be assessed (making explicit implicit hypotheses, showing well non-well posed questions), decomposition as a way to simplify assessments (use of influence diagrams, event trees and uncertainty propagation techniques) and adequate evaluation of different evidences in order to assess probabilities (use of Bayes' theorem and concepts of independence and reliability of information sources).

The last part of the training session is dedicated to explain Knowledge biases to the experts in order to teach them to provide more reliable opinions, i.e.: representativity, availability and anchor and adjustment. Experts should be informed about the hazard of being overconfident. A calibration exercise could be appropriate. The whole training session should not take more than one morning.

Tasks definition

This step is done through an interactive session of the project team and the experts. The issue at hand is to explain them in a detailed way as well as the questions to be assessed, and to make a schedule of the activities to be developed by each expert. All the work developed by the project team during the *Definition of the Problem* phase should be used now. The session should start with a presentation given by the generalist about the parameters to be assessed, including all relevant sources of information previously identified. Experts should provide their own view of the problem and the definition of the parameters, pointing out, if needed, further information sources, computations to be done, etc. The result of this session would eventually be a refined definition of the parameters under study. Common definitions to all the experts should be agreed.

The second step in this meeting is to study the possible ways to decompose each parameter. The project team should provide a seminal decomposition that should be discussed with the experts. The objective is to help the experts to develop their own decompositions. Decompositions could be quite different from one expert to another one. Each expert will have to assess uncertainties of variables at the lowest levels. The analysts will usually do the aggregation of all uncertainties. This is the point to introduce propagation of uncertainties concepts to the experts and to let them know all the potential variety of tools that the analysts could provide them to pre-process and post-process probabilistic runs of computer codes, or of the simple decomposition model developed by experts.

Individual experts' work

Experts develop their analysis during this phase, according to the schedule agreed in the previous step. Each expert will write by the end of this period a report summarising the main hypothesis and procedures used during his/her work, the conclusions achieved and, if he/she wishes, a preliminary assessment of uncertainties. Whenever needed during this period, the team project should be available to each expert in order to provide statistical support or to solve any doubt about the parameters to be assessed.

At the end of this phase, the team project organises a meeting with all the experts. Each expert presents his/her work and the conclusions achieved. This meeting allows each expert to get some hints about alternative ways to tackle the problem.

Elicitation of experts' opinions

The elicitation of each expert opinion's is individual and should be done in a quiet environment, if possible without interruptions. It is convenient to have the presence of an analyst and a generalist, in

addition to the expert. In a systematic way, the analyst gets the opinion of the expert for each parameter, asking for supporting reasons whenever necessary. The role of the generalist in this session is to provide additional information when needed, to provide general support and to audit the session in order to avoid irregularities (bias induction, etc.). Whenever needed, the analyst could ask questions in a different way to check potential inconsistencies. The session must be recorded as much as possible (tape recorders, video or extensive hand annotations).

The techniques used to help the expert when assessing uncertainties are quite standard: quantile assessment for continuous variables and probability estimations for discrete variables (direct or indirect methods); in the case of experts with some skills in probability other techniques like direct parameter assessment or drawings are acceptable.

Analysis and aggregation of results

Assessments provided by experts are studied in this phase. The objective is to check that there is no important bias and the logic correctness of their rationale. If biases and logic faults are not present in expert's assessments, next step is to check if individual opinions may be aggregated to get a unique distribution for each parameter.

Before aggregating individual distributions the overlap between distributions of different experts should be assessed. If the distributions do not overlap the experts disagree. In that case aggregation should be avoided. Under these circumstances a reconciliation session could be of help. An analyst should lead the session and should organise it according to the following steps:

1. Exposition of different opinions.
2. Identification of differences.
3. Discussion about the reasons for each original assessment.
4. Discussion about the different sources of information used.
5. Re-elaboration of individual opinions in posterior elicitation sessions or joint assessment (through consensus) of a common distribution, if agreed by experts.

In the case that a consensus distribution is obtained, that is the final step (before documentation). If further elicitation sessions are needed, the consistency of the opinions is checked again and aggregation is done if acceptable overlap is achieved. Otherwise, the project team should choose what opinions could be aggregated as main opinion of the group (after aggregation), and what opinions should be left as an alternative to perform sensitivity analysis. The main strategies for aggregation are the following ones:

1. Linear combination.
2. Log-linear combination.
3. Bayesian combination.

Documentation

Documentation of the application must be as complete as possible, including results and description of the ways to obtain them. The contents of the documentation should follow the order of application of the procedure, recording, in each step, *what* has been done, *why* it has been done, *how* it has been done and *Who* has done it. In order to achieve this degree of documentation, a schedule of standardised documentation activities should be made for each phase. It should always be completely clear to the reader what a result assessed by an expert is as well as the outcome of an aggregation, sensitivity analysis or any other analysis not provided explicitly by an expert.

5. Propagation of uncertainties

Whenever the system model is available and the distributions of the input parameters have been derived, the next step in the PA study is to propagate uncertainties in order to get information about the distribution of the output variables. Analytical uncertainty propagation methods can only be used for very simple systems with very few parameters. In more complex cases other methods need to be adopted. The most suited method, and in fact the most used one is the Monte Carlo method.

5.1. The Monte Carlo method

The Monte Carlo method consists in sampling at random the vector of input parameters, running the system model computer code for each sample of that vector and getting a sample of the vector of output variables. Later on, the characteristics of the output variables may be estimated using the output samples obtained. One of the advantages of using the Monte Carlo method is that all statistical standard methods we need to estimate the output variables distributions and to test any hypothesis may be used. This makes it the most straightforward and powerful method available in the scientific literature to deal with uncertainty propagation in complex models. This method is valid for models that have static and also dynamic outputs. It is adequate for working with discrete and continuous inputs and outputs, and the implementation of computational algorithms required has no fundamental complexity.

The Monte Carlo maps the input space into the output space point by point. In order to see this, let us consider a very simple model: $Y=X_1+X_2$. Suppose X_1 and X_2 follow independent uniform distributions both of them defined in the interval [0,1]. For this simple model an analytical propagation of uncertainties is feasible and the output Y follows a triangular distribution defined in the interval [0,2] and whose mode, mean and median are 1. This propagation may be done via Monte Carlo. First, a sample of size 100 is taken in the input space (see Figure 27). For each point shown in Figure 27, the value of the output is then computed. An empirical cumulative distribution function is built from the 100 values obtained (see Figure 28). For the sake of comparison the actual CDF of Y has also been drawn.

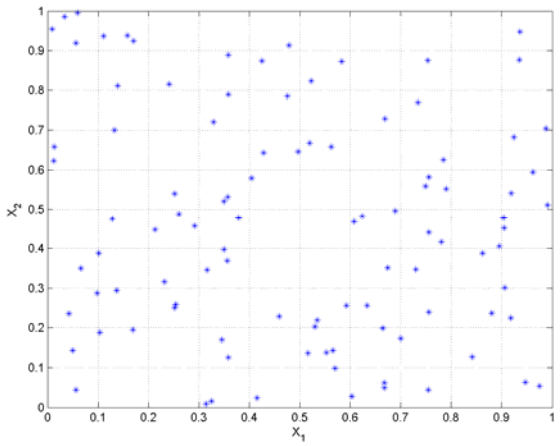


Figure 27: Simple random sample of size 100 of two random variables uniformly distributed in the region [0,1]x[0,1].

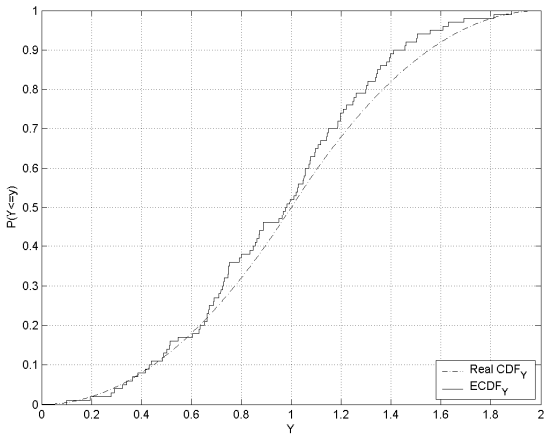


Figure 28: ECDF obtained from a simple random sample of size 100 of $Y=X_1+X_2$ and the theoretical CDF of Y.

Monte Carlo may also be seen as a numerical integration method. In the same example, let us consider that we are primarily interested in the estimation of the mean of Y. This means that we are trying to estimate

$$\mu_Y = \int_{[0,1]\times[0,1]} (X_1 + X_2) dx_1 dx_2 . \tag{5.1}$$

One of the possible approximations to compute this integral is to take the sample considered in Figure 27 and Figure 28 and to calculate the arithmetic mean

$$\hat{\mu}_Y = \frac{1}{100} \sum_{i=1}^{i=100} \left(x_{1i} + x_{2i} \right). \tag{5.2}$$

It is important to remark that the standard deviation of this estimator is

$$\sigma_{\hat{\mu}_Y} = \sigma_Y / \sqrt{n} , \tag{5.3}$$

where σ_Y is the standard deviation of the output Y. Figure 29, which shows the histogram obtained from 50 simple random samples of size 100, similar to the one shown on Figure 27 and Figure 28. In this plot we can see that the range of $\hat{\mu}_Y$ is roughly 0.2, which means it represents one tenth of the range of Y (the range of this triangular distribution is 2). Figure 30 shows the corresponding ECDFs. It is important to remark that the standard error of $\hat{\mu}_Y$ does not depend on the dimension p of the space where the integral is computed, and that consequently the Monte Carlo method does not suffer from the curse of dimensionality. [Metropolis 49] is the seminal paper about Monte Carlo, where many interesting suggestions are made about its applicability.

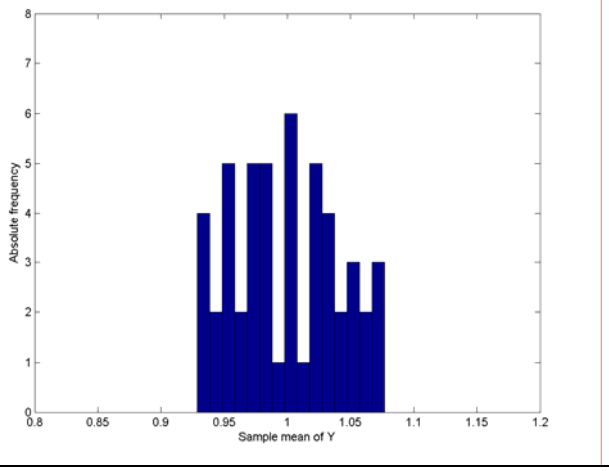


Figure 29: Histogram of the sample means obtained from 50 simple random samples of size 100 obtained via Monte Carlo simulation.

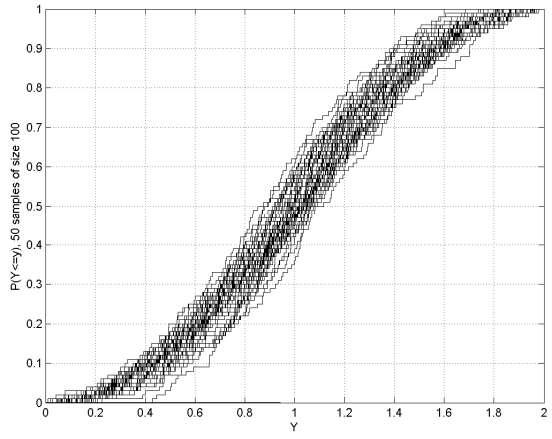


Figure 30: ECDFs obtained from 50 simple random samples of size 100 obtained via Monte Carlo simulation.

5.2. Variance reduction techniques

The computational time to perform a Monte Carlo analysis depends on the number of simulations and cost for each simulation. The computational time for complex problems with a large number of simulations often become prohibitive. The cost for each simulation can be reduced by simplifying the mathematical description of the problem. A second alternative is to reduce the number of simulations

compared to standard random sampling without scarifying the precision and confidence intervals of the outputs. Such techniques are referred to as Variance Reduction Techniques. Main techniques are Latin Hypercube sampling (LHS), stratified sampling, control variates, importance sampling and antithetic variates ([Rubinstein 81], [Fishman 96] and [Robert 04]). In the following pages we discuss about most relevant variance reduction techniques.

Stratified sampling

Input parameters may vary considerably. By stratification the population is sub grouped into relatively homogenous subgroups. The sampling is then performed for each of the strata. The strata must be mutually exclusive and collectively exhaustive.
Stratified sampling is based on the fact that the variance of any random variable, once it has been divided in strata, may be decomposed into two contributions: the variability within each stratum and the variability between different strata, which means

$$\sigma^2 = \sum_{i=1}^{h} \omega_i \cdot \sigma_i^{\,2} + \sum_{i=1}^{h} \omega_i \cdot (\mu_i - \mu)^2 \,, \tag{5.4}$$

where the first summand represents the variability within the h considered strata and the second one represents the variability between different strata. ω_i stands for the probability of stratum i, and μ_i and σ_i stand for the mean and the standard deviation of the (output) variable Y also in stratum i. If the sampling of each observation is restricted to a given stratum, its variability will be the variability of that stratum (σ_i) rather than the whole variability of Y (σ). The estimate for the mean of Y under stratified sampling is

$$\hat{\mu}_S = \sum_{i=1}^{h} \omega_i \hat{\mu}_i = \sum_{i=1}^{h} \omega_i \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad , \tag{5.5}$$

where $\hat{\mu}_i$ is the estimate of the mean of Y in stratum i, which normally is computed as the average of the values of Y obtained in that stratum (y_{ij}), as shown in (5.5). n_i is the sample size in stratum i and the whole sample size is $n=n_1+\ldots+n_h$. This estimator is an unbiased estimator of Y's mean. It may be easily demonstrated that the variance of this estimator for a given number of samples is reduced with respect to the estimator provided by simple random sampling according to

$$Var(\hat{\mu}_S) = Var(\hat{\mu}) - \frac{1}{n} \sum_{i=1}^{h} \omega_i (\mu_i - \mu)^2 \quad , \tag{5.6}$$

which means that the larger the differences between the means of the different strata the larger the decrease in the variance of the stratified sampling estimator.

The main problem affecting stratified sampling is that ideally what should be stratified is the output space, so that the second term on the right hand side of (5.6) would be large and so it would be profitable to stratify. Unfortunately, what can be easily stratified is the input sample space, which doesn't mean that the corresponding stratification in the output space will be so good. Under those circumstances, when large overlaps between different strata happen, the benefit from stratifying would not be so important, though some benefit will always be obtained according to (5.6).

Once the sample size has been chosen, there are two problems to be solved: 1) how to create the strata and 2) the sample size within each stratum. There is no clear rule to partition the input sample space. When no additional information is available about the system model, the most common strategy is to build a net of hypercubes via Cartesian product of the stratification performed in each input variable.

When some information is available, it can be used for creating the stratification. In general, there are two ways to get information about the model: Studying the equations of the model and getting a small size sample. The study of the equations of the model may provide information on the relation between inputs and outputs in the model and on the importance of combinations of specific sets of inputs, see section 5.3. A small size sample could be obtained via simple random sampling and it could be used to perform Sensitivity Analysis (SA). The use of SA techniques could help identifying the most relevant input variables; stratification could be performed only on these relevant input variables

Regarding the sample size per stratum, there are several options. The first option is to take proportional sampling, which means that the sample size in each stratum is proportional to the probability of the stratum: $n_i=n\omega_i$. Further improvement may always be achieved ([McKay 79], [Fishman 96]) if the sample space is further stratified to getting as many strata as samples (one observation per stratum). In that case the reduction in the variance of the estimator with respect to simple random sampling is

$$Var(\hat{\mu}_S) = Var(\hat{\mu}_R) - \frac{1}{n^2}\sum_{i=1}^{n}(\mu_i - \mu)^2 \quad . \tag{5.7}$$

Not only the mean of the output is more accurately estimated when stratified sampling is applied, but its distribution is better estimated due to the evenness in the sampling all over the sampling space, no region is either over-sampled or under-sampled. Figure 31 provides an idea about the way to get a stratified sample of size 9 with one observation per stratum (each stratum has probability *1/9*) in a 2-D input sample space. Figure 32 provides the same information when the stratification is done on only one of the input variables (*X*).

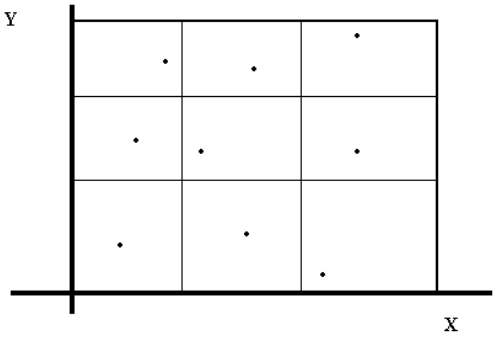


Figure 31: Stratified sample with nine observations for two variables; one observation per stratum, probability of each stratum 1/9.

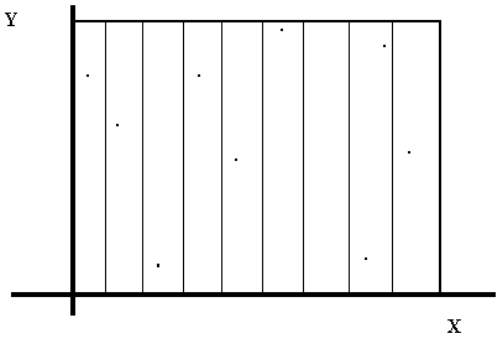


Figure 32: Stratified sample with nine observations for two variables; one observation per stratum, probability of each stratum 1/9.

Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling (LHS) is a cost-effective and reliable extension of stratified sampling, designed to generate collections of parameter values from multivariate distributions. In order to get a sample of size *n*, the procedure is the following one:

1. a stratified sample is obtained for each input variable (n strata with probability 1/n each one and a sample of size 1 per stratum),
2. get a permutation of each one of the samples of each input variable

3. combine the first observations of all the variables (after permutation) to get the first observation of the input vector, combine the second observations of all the variables (after permutation) to get the second observation of the input vector and so on.

The procedure above is generally complemented by techniques to fill the input space in an optimal way, for instance by maximizing the minimum distance between the samples points. [McKay 79] shows LHS produces unbiased estimators for the mean and the CDF of the output. They also demonstrate that a sufficient condition to get an estimation error for the sample mean and the CDF smaller that in the case of random sample is that the model has to be monotonic in all its input variables. [Stein 87] proved some asymptotic properties of LHS under general conditions: the variance of the estimators provided for the mean and the CDF are smaller (asymptotically) than the ones obtained under simple random sampling, with the degree of variance reduction depending on the additivity of the model. The estimates do also follow, asymptotically, a normal distribution. [Iman 82] developed a method to induce rank correlation between input variables sampled under this scheme and [Stein 87] introduced a method to induce correlations between input variables. Figure 33 shows the way to generate a sample of size 5 through this method for a bivariate random vector.

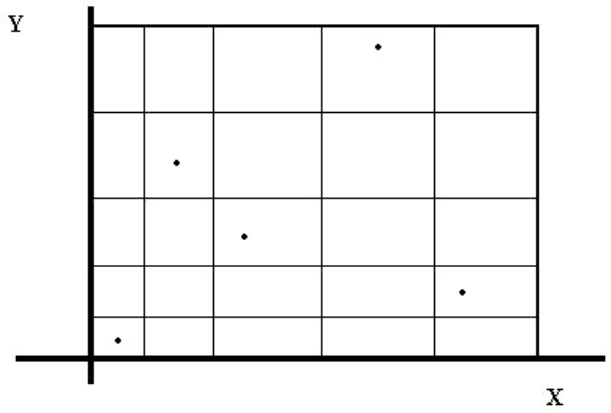


Figure 33: LHS sample of size 5 for two variables. Each stratum has the same probability

Importance sampling

Importance sampling was designed to estimate more accurately the mean of a random output variable rather than its entire distribution. This method is based on substituting the variable under study by another one that has the same mean but a smaller variance. In order to get this, let us observe that the mean of the output variable in the system model (Y) may be written in this way

$$\mu = \int_{S_x} y(\mathbf{x}) \cdot f(\mathbf{x}) \cdot d\mathbf{x} = \int_{S_x} y(\mathbf{x}) \cdot \frac{f(\mathbf{x})}{g(\mathbf{x})} \cdot g(\mathbf{x}) \cdot d\mathbf{x} \quad , \qquad \textbf{(5.8)}$$

so that the variable whose mean is actually studied is $Y'=Y'(\mathbf{x})=Y(\mathbf{x})\cdot f(\mathbf{x})/g(\mathbf{x})$, where the only requirement on the function $g(\mathbf{x})$ is that it must be a PDF. Under these conditions, the variance of Y' is

$$Var(Y') = \int_{S_x} (y(\mathbf{x}) \cdot \frac{f(\mathbf{x})}{g(\mathbf{x})} - \mu_y)^2 \cdot g(\mathbf{x}) \cdot d\mathbf{x} \quad , \qquad \textbf{(5.9)}$$

The more constant (stable over the values of $\boldsymbol{X}$) the function $Y(\mathbf{x}) \cdot f(\mathbf{x})/g(\mathbf{x})$ is, the smaller is the variance (in (5.9)). In order to reduce the estimation error we need to sample more with increasing probability of a region of the input space and also if the values of the output variable in that region are large. It is important to notice that, since by construction the means of Y and Y' are the same, we will estimate the mean of Y' because we are able to do it in a more accurate way, and this estimation will also be valid for Y. Figure 34 shows the idea behind importance sampling.

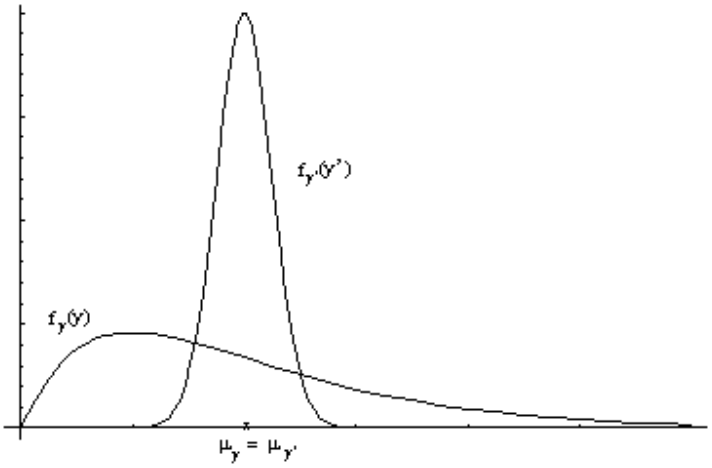


Figure 34: Intuitive idea behind importance sampling

Control variates

This technique is based on decomposing the output random variable Y as a sum of two ancillary random variables Y' and Y'' in such a way that Y' should have a mean analytically and easily computable, or at least with a well known dependence on the vector of input parameters X (so that its mean could be computed with the needed accuracy at a low cost), and Y'' should have a small variance. Under these conditions, the mean may be split up as

$$\mu_y = \int_{S_x} y(\mathbf{x}) \cdot f(\mathbf{x}) \cdot d\mathbf{x} = \int_{S_x} y'(\mathbf{x}) \cdot f(\mathbf{x}) \cdot d\mathbf{x} + \int_{S_x} (y(\mathbf{x}) - y'(\mathbf{x})) \cdot f(\mathbf{x}) \cdot d\mathbf{x} \ , \tag{5.10}$$

where $y''(x) = y(x) - y'(x)$ (Figure 35). Again, as in the case of importance sampling and stratified sampling, we need additional information to find Y'. If no such theoretical information is available, the most straightforward way to get it is using a previous small size sample. That sample may be used, for instance, to build a response surface (see [Myers 02]) that captures the main characteristics of the functional dependence of Y over X. The response surface obtained would be Y' (also represented as $y'(x)$ in this text). On one side Y' will usually be a polynomial that may be used to propagate uncertainty analytically or computationally using huge sample sizes, estimating the first integral on the right hand side of (5.10) with no or negligible error. On the other side, if the quality of this response surface is good, $y(x) - y'(x)$ would have small values for all values of input vector x, so that the last integral in (5.10) would be the only one introducing relevant error in the estimation of the mean, but much smaller that the one introduced by normal random sampling.

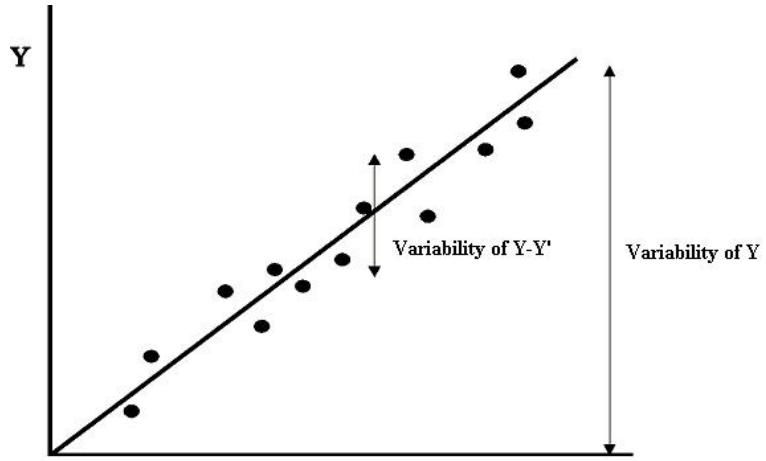


Figure 35 : Intuitive idea behind control variates sampling technique.

5.3. Dimension reduction

Let us consider a system of functional equations where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are the dependent or output variables and $\boldsymbol{X} = (X_1, \ldots, X_m)$ are the independent variables (e.g., space coordinates and time). Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ be the parameters of the system, that is, coefficients of the differential equations and of the initial and boundary conditions. The solutions of the system are $Y_j = I_j(\boldsymbol{X}; \boldsymbol{\theta})$.

In physics, one speaks of similarity between two problems when one can transform one problem into the other by a change of scale in the variables. It is shown that this is possible when a set of dimensionless numbers (in mathematical terms, we shall speak instead of invariant functions), which are functions of the parameters $\boldsymbol{\theta}$; coincide in both problems. A classical example is the Reynolds number in fluid mechanics. The dimension of the parameter space, originally p, can thus be reduced to the number of dimensionless quantities that define the system of functional equations. This problem is referred to in the literature as dimensional analysis, and though in many physics and engineering works it is formulated in terms of physical magnitudes and dimensions ([Buckingham 14]; [Langhaar 51]; [Palacios 64]; [Szirtes 98]). A more abstract, mathematical, and hence physics independent language, is preferable when dealing with propagation of uncertainties, such as in [Moran 71].

[Moran 72] generalized dimensional analysis consists of finding a set of linear transformations:

$$\begin{aligned} Y'_j &= K_j Y_j \,(j = 1,\ldots, n \geq 1) \\ X'_k &= K_{n+k} X_k \,(k = 1,\ldots, m \geq 1) \\ \theta'_e &= K_{n+m+e} \theta_e \,(e = 1,\ldots, p \geq 1) \end{aligned} \tag{5.11}$$

of the $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\theta}$, where the $K_j, j = 1, \ldots, n + m + p$ are constants, such that the system of functional equations is invariant under the transformations, that is, $Y_j = I_j(\boldsymbol{X}; \boldsymbol{\theta})$ transforms to $Y'_j = I'_j(\boldsymbol{X'}; \boldsymbol{\theta'})$; where $\boldsymbol{X'} = X_1', \ldots, X'_m$ and $\boldsymbol{\theta'} = (\theta'_1, \ldots, \theta'_p)$. We note that the prime symbol stands for variable transformation and not for array transposition. A more general class of transformations could have been used, but we are restricted here to linear transformations (scale changes) because they have proved useful in many physical problems, while maintaining mathematical simplicity and a clear physical interpretation.

After introducing the transformations or scale changes into the system equations and boundary and initial conditions, and imposing the condition of invariance (the system equations maintain the same form before and after the transformation; $Y_j = I_j(\boldsymbol{X}; \boldsymbol{\theta}) \Leftrightarrow Y'_j = I'_j(\boldsymbol{X'}; \boldsymbol{\theta'})$), there appear restrictions linking the values of the $K_i, i = 1, \ldots, n + m + p$. In most cases, the restrictions will reduce their

degrees of freedom. So if initially there are n+m+p transformation constants K_i and q restrictions, there will finally be $r = n+m+p-q$ degrees of freedom for the K_i. Then, the transformations can be defined in terms of a reduced set of constants, which are called A_j, $j=1,...,r$, and the set of transformations may me rewritten as

$$\begin{aligned} Y'_j &= A_1^{a_{j1}} \ldots A_r^{a_{jr}} Y_j \; (j = 1,\ldots, n \geq 1) \\ X'_k &= A_1^{b_{k1}} \ldots A_r^{b_{kr}} X_k \; (k = 1,\ldots, m \geq 1) \\ \theta'_e &= A_1^{c_{e1}} \ldots A_r^{c_{er}} \theta_e \; (e = 1,\ldots, p \geq 1) \end{aligned} \tag{5.12}$$

where the a_{ji}, b_{kl} and c_{et} are exponents. In fact each restriction defines an invariant function or dimensionless number

$$\pi = Y_1^{\alpha_1} \ldots Y_1^{\alpha_n} X_1^{\beta_1} \ldots X_1^{\beta_n} \theta_1^{\gamma_1} \ldots \theta_1^{\gamma_p} \tag{5.13}$$

where the α_i, β_j, γ_k are also exponents, in such a way that (see [Moran 72]) the system of functional equations can be expressed in terms of these invariant functions, instead of in terms of the original and larger set formed by $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\theta}$. The calculation of the invariants and of the expression of the system model in terms of the invariants is formalized in the theorems of [Moran 72]; see also appendix A of [Mira 04] for details.

Usually, the reduction of dimension is in the space of input parameters and input variables ($\boldsymbol{X}$, $\boldsymbol{\theta}$), only very infrequently is the reduction performed in the space of output variables. Even when a reduction of dimension is obtained in the space of input parameters, it does not necessarily mean that this produces a benefit in the propagation of uncertainties. It is possible that the reduction of dimension happens in the part of the space of input parameters that is not affected by uncertainty (known constants); in that case no improvement is obtained. Moreover, in order to get some benefit, variance reduction techniques have to be applied in combination with dimension reduction. If an effective dimension reduction is obtained, using simple random sampling on this space doesn't lead to a net decrease in the variance of the estimators of the outputs; a simple random sample of the input space produce a simple random sample of the output space independently of the dimension of the equivalent input sampled space.

[Mira 04] describe an application of dimension reduction obtained via dimensional analysis for the propagation of uncertainties of a simplified HLW repository. In this application, the original space of input parameters and input variables has dimension 7 and the transformed one 4. Nevertheless, the real reduction obtained is from 3 to 2 since only two input parameters and one input variable are affected by uncertainty and these inputs are concentrated in only 2 invariants in the transformed input space. Mira and his colleagues compare in their work four sampling techniques: simple random sampling, LHS and stratified sampling in the original 3-D input space and stratified sampling in the 2-D transformed input space. For this comparison 60 samples of size 64 were used. Figure 36 shows the means of the means for the flow of ^{129}I getting into the biosphere at different times. All techniques produce unbiased results. Figure 37 show the standard deviations of the means for the same case and illustrates the improvement that is obtained when combining dimension reduction and stratified sampling with respect to the other techniques applied on the original input space. The results for LHS in the 2-D space shouldn't be taken into account since it shouldn't be called LHS the way this sampling scheme was actually applied in this test case.

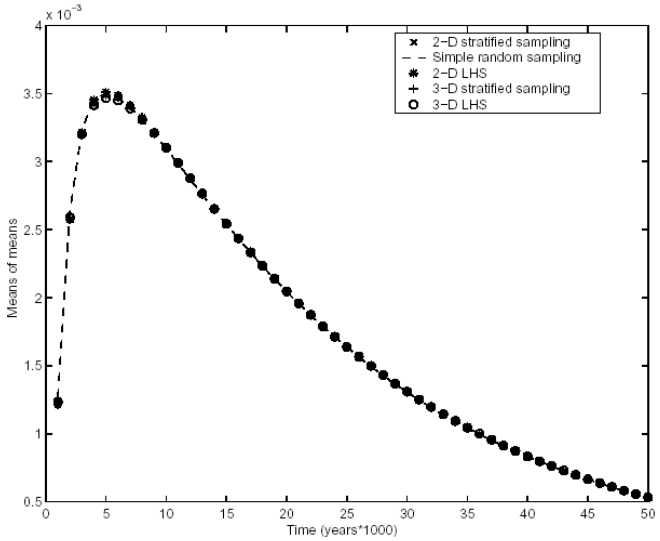


Figure 36: Evolution over time of the mean of the means for different sampling schemes with (dimensions reduction (2-D) curves) and without (no dimension reductions (3-D) or simple random curves) input space dimension reduction obtained via Dimensional Analysis.

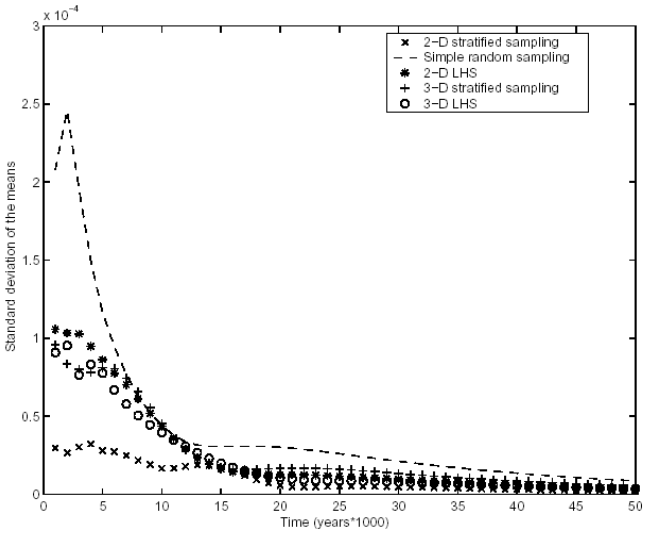


Figure 37: Evolution over time of the standard deviation of the means for different sampling schemes (2-D curves) and without (3-D or simple random curves) input space dimension reduction obtained via Dimensional Analysis.

Dimension reduction in the input space may also be obtained in a more immediate, less sophisticated way, referred to as trivial reductions of dimension. It frequently occurs that in the differential equations which describe the behavior of the system, some coefficients appear as elementary functions of a number of, let us say, original coefficients, such that, either due to physical reasons or because of reasons related the way experts address the problem, the uncertainty has been expressed in terms of probability distributions for the original coefficients. [Bolado 04] showed that, as in the case of dimension reduction obtained via dimensional analysis, smaller errors are made in the estimation of the outputs when trivial reductions of dimension are combined with variance reduction techniques such as stratified sampling.

5.4. Choice of the number of samples (Wilks)

The larger the sample size, the more information the sample contains. Nevertheless, increasing the sample size means increasing the computational cost. So there is balance between the information we can obtain and its cost. In order to set sample sizes we will consider two concepts: conservative estimates of quantiles and tolerance intervals.

Quantile estimation is a very important issue when dealing with outputs of PA codes since, in many occasions, safety limits are based on quantile estimation. This is the case, for example, when a safety limit is set in the following terms: In order to be acceptable, the repository should not produce, at any time in the future, an annual individual dose higher than D_L $mSv \cdot y^{-1}$ with a probability higher than 0.05. This means that the percentile 95% of the output variable 'annual individual dose' should be lower than D_L.

In the case of a real-valued rv Y, quantile estimation means determining the level y such that the likelihood that Y takes a value lower than y is some prescribed value. Using the CDF of Y, $F(y) = P(Y \leq y)$, we seek an estimation of the α-quantile y_a defined by $F(y_a)= \alpha$.

In the following we will describe how to set minimum sample sizes to achieve given quantities of information. Most of the ideas used come from the theory of order statistics. The references for this section are : [Cannamela 07], [David 03], [Guba 03], [Makai 06], [Nutt 04], [Nutt 05], [Orechwa 05], [Wallis 03], [Wallis 06], [Wilks 41].

5.4.1. Empirical estimator

Let $\hat{F}_{EE}(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \leq y}$ be the empirical estimator of the CDF, where $1_{Y_i \leq y}$ is the indicator function (one for $Y_i \leq y$ and zero for $Y_i > y$). This leads to the following estimator of the α- quantile

$$\hat{Y}_{\alpha,n} = \inf\{y, \hat{F}_{EE}(y) > \alpha\} = Y_{(\lceil \alpha n \rceil)}. \qquad (5.14)$$

The properties of this estimator are given in Annex 1, 8.1. The variance of this estimator is large. Moreover, from the asymptotical law, we obtain $P(\hat{Y}_{\alpha,n} \geq y_\alpha) \approx 0.5$, which comes from the fact that this is an asymptotically unbiased estimator of the α- quantile. So, for sufficiently large sample sizes (we never know if the sample size actually used is large enough), roughly 50% of the times our estimate will be larger and 50% of the times it will be smaller than the actual α- quantile. When working in the area of safety, we are usually interested in estimating extreme (high) quantiles, such as the 95%, the 99%, etc. In those cases, it could be advisable to be more confident that our estimate is above the actual quantile. Then we would be interested in a conservative estimator of the α- quantile like Wilks' introduced below.

5.4.2. Wilks estimator

In order to set a conservative estimator for the quantile of interest (y_α), we should decide how confident we want to be that our estimator exceeds the real quantile. The confidence level β would typically be either 0.95 or 0.99. The estimator considered is the one that fulfils

$$\hat{Y}_{\alpha,n} = Y_{(\lceil \alpha n \rceil + s)}, \text{ such that } P(Y_{(\lceil \alpha n \rceil + s)} \geq y_\alpha) = \beta \ , \tag{5.15}$$

where s ≥ 1. This estimator, based on order statistics, is referred to as Wilks estimator, and is based on the probabilistic distribution of the number of times a sample of the rv exceeds a certain threshold. Let us rename it as $Y_{(n-r+1)}$. For each couple (*n*,*r*) we will get a given value $P(Y_{(n-r+1)} > y_\alpha)$, which is (see Annex 1, 8.2 for more details)

$$\sum_{j=0}^{n-r} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} \tag{5.16}$$

For a fixed *r*, we may (numerically) compute the smallest value of *n* needed to make this expression (5.16) larger than or equal to β (for fixed values of β and α).

Interpretation: *n* is the number of simulations required for the r^{th} largest value of the ordered sequence of outputs to exceed the α- quantile with a prescribed confidence level β.

Example: For α = β = 95%, we have the following couples :
(*r* = 1, *n* = 59) ; (*r* = 2 , *n* = 93) ; (*r* = 3, *n* = 124) … (*r* = 39 , *n* = 991).

For *r*=1, the previous formula becomes

$$\beta = 1 - \alpha^n \ . \tag{5.17}$$

Remark: The variance of this estimator is even larger than the one of the empirical estimator (see Figure 38).

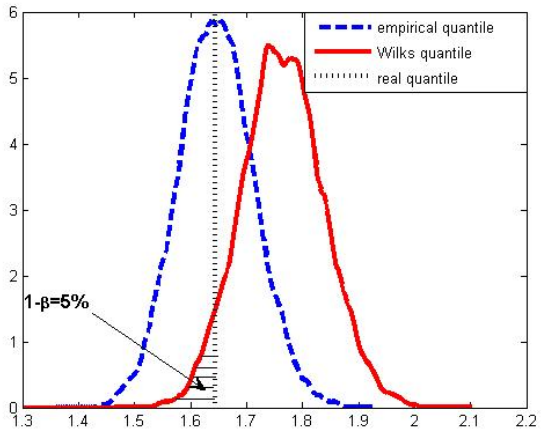


Figure 38: Comparison between distributions of the empirical and the Wilks estimators, for samples from a normal distribution (the real 95% quantile for the normal distribution is 1.6449). For this example, the variance of the empirical estimator is 0.0045, while for the Wilks estimator is 0.0053.

In conclusion, the problem of quantile estimation can be solved by Monte Carlo or LHS techniques, but the the estimates are imprecise (i.e. with large variance) if the number of runs is often considered "reasonable" (100 – 1000 runs).

5.4.3. Tolerance interval

An alternative solution to set a sample size is to estimate a tolerance interval rather than a percentile ([Guba 03], [David 03]). A tolerance interval has random bounds, denoted by L (lower) and U (upper)

and the requirement for this interval is that it should contain at least a proportion γ of the population, with probability β (with prescribed γ and β). Hence we seek L and U such that

$$P\left[\int_{L}^{U} f(y)dy \geq \gamma\right] = \beta \tag{5.18}$$

where *f* is the (unknown) underlying PDF.

It has been shown (see for instance [David 03]) that the left hand side of equation (5.18)is independent of *f* if and only if the bounds L and U are order statistics (i.e. $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$). To see the necessary condition, let $L = Y_{(r)}$ and $U = Y_{(s)}$, $r < s$ (where $Y_{(0)} = -\infty$ and $Y_{(n+1)} = +\infty$), and then the equation (5.18) may be written as $P[F(Y_{(s)}) - F(Y_{(r)}) \geq \gamma] = \beta$. The quantities $F(Y_{(s)}), F(Y_{(r)})$ are order statistics for a uniform distribution in [0,1]. The distribution of the range (here the range is $F(Y_{(s)}) - F(Y_{(r)})$) of order statistics is known for uniform distributions and is given by [David 03]:

$$P[F(Y_{(s)}) - F(Y_{(r)}) \geq \gamma] = 1 - I_{\gamma}(s-r, n-s+r+1) = \beta\,, \tag{5.19}$$

where $I_{\gamma}(j,k)$ is the incomplete beta function[1]. Equation (5.18) and (5.19) are not satisfied exactly, but some values of *r* and *s* may be chosen such that

$$P\left[\int_{L}^{U} f(y)dy \geq \gamma\right] \geq \beta\,. \tag{5.20}$$

Application 1: When a minimum sample size is sought, we are focusing our attention on what happens when the sample maximum and minimum are selected to estimate a tolerance interval (standard bilateral tolerance interval: $L = Y_{(1)}$, $U = Y_{(n)}$), we obtain from (5.19) the following value for β:

$$\beta = 1 - \gamma^{n} - (n-1)(1-\gamma)\gamma^{n-1}, \tag{5.21}$$

which may be solved numerically for *n* and the result rounded to the next highest integer.

The following numerical example is given in [Guba 03]. For γ=0.953 and β=0.95 we obtain n=100, and we should select the lowest value as *L* and the largest one as *U*.

[1] Which is defined by $I_{\gamma}(j,k) = \int_{0}^{\gamma} \frac{u^{j-1}(1-u)^{k-1}}{B(j,k)} du$, $B(j,k) = \frac{(j-1)!(k-1)!}{(j+k-2)!}$.

n	γ values		
	β=0.90	β=0.95	β=0.99
10	0.66315	0.60584	0.49565
20	0.81904	0.78389	0.71127
30	0.87643	0.85141	0.79845
40	0.9062	0.88682	0.84528
50	0.92443	0.9086	0.87448
60	0.93671	0.92336	0.89442
70	0.94557	0.93402	0.9089
80	0.95225	0.94207	0.91989
90	0.95747	0.94837	0.92851
100	0.96166	0.95344	0.93554
125	0.96924	0.96262	0.94813
150	0.97432	0.96877	0.95658
175	0.97796	0.97318	0.96268
200	0.98069	0.9765	0.96736
225	0.98282	0.97909	0.97087
250	0.98453	0.98118	0.97375
275	0.98593	0.98287	0.97618
300	0.9871	0.98429	0.97809

Table 3: Values of γ for the standard tolerance interval for diferent values of n and of β (from [Guba 03]).

Application 2: For the standard **unilateral tolerance interval** case, when $L = Y_{(0)}$, $U = Y_{(n)}$, we obtain the following value for β:
$\beta = 1 - \gamma^n$, which is exactly the same result obtained in the case of the Wilks quantile estimator (and again we get n=59, for $\beta = 95\%, \gamma = 95\%$).

5.5. Metamodels

The numerical codes used in safety demonstration for NPP or in PA studies are very time-consuming, and this makes them difficult to use for probabilistic studies, where thousands of simulations are needed to produce an accurate result. An alternative is to simplify the numerical code by *approximating* it. The approximation of the numerical code is called *metamodel* (or *response surface* or *surrogate model*). The metamodel is hence used instead of the numerical code for the probabilistic studies such as uncertainty propagation and sensitivity analysis. Different techniques, developed initially for modelling the real laboratory experiments (see [Box 87] and [Myers 02] for example) have been adapted to *numerical experiments* (where a numerical experiment is a run of the numerical code), see for instance [Sacks 89a], [Sacks 89b], [Welch 92], [Santner 03].

The response surface methodology consists in approximating the numerical code by an appropriate empirical (and simple) model of the form $y(\boldsymbol{x}) = f(x_1, \ldots, x_d) + \varepsilon$ (where y is the output (or the response variable), $\boldsymbol{x} = (x_1, \ldots, x_d)$ are the inputs (or the predictor variables, or the design variables) and ε is the error (or the residual)). This simple model has to be identified and fitted from *experimental* data (i.e. results of numerical simulations performed with the original numerical code). To obtain the final metamodel three steps have to be performed:

- select a design of experiments (i.e. the points for which the numerical experiments have to be performed)
- use some modeling techniques (such as linear models (LM) [Myers 02], generalized additive models (GAM) [Hastie 90], multivariate additive regression spline (MARS) [Friedman 91]

and polychotomous regression based on MARS (PolyMARS) [Kooperberg 97], kriging [Santner 03], neural networks,…)
- use optimization techniques to fit the parameters of these models.

The final purpose is to obtain a cost-effective model with good predictive capacities.

5.5.1. Design of experiments

The design of experiments is a discipline which selects the experiments to be performed by simultaneously varying the input variables, in order to extract as much information as possible, with a low cost (low CPU time). Moreover the number of experiments to be performed has to be large enough to estimate the metamodel's parameters.

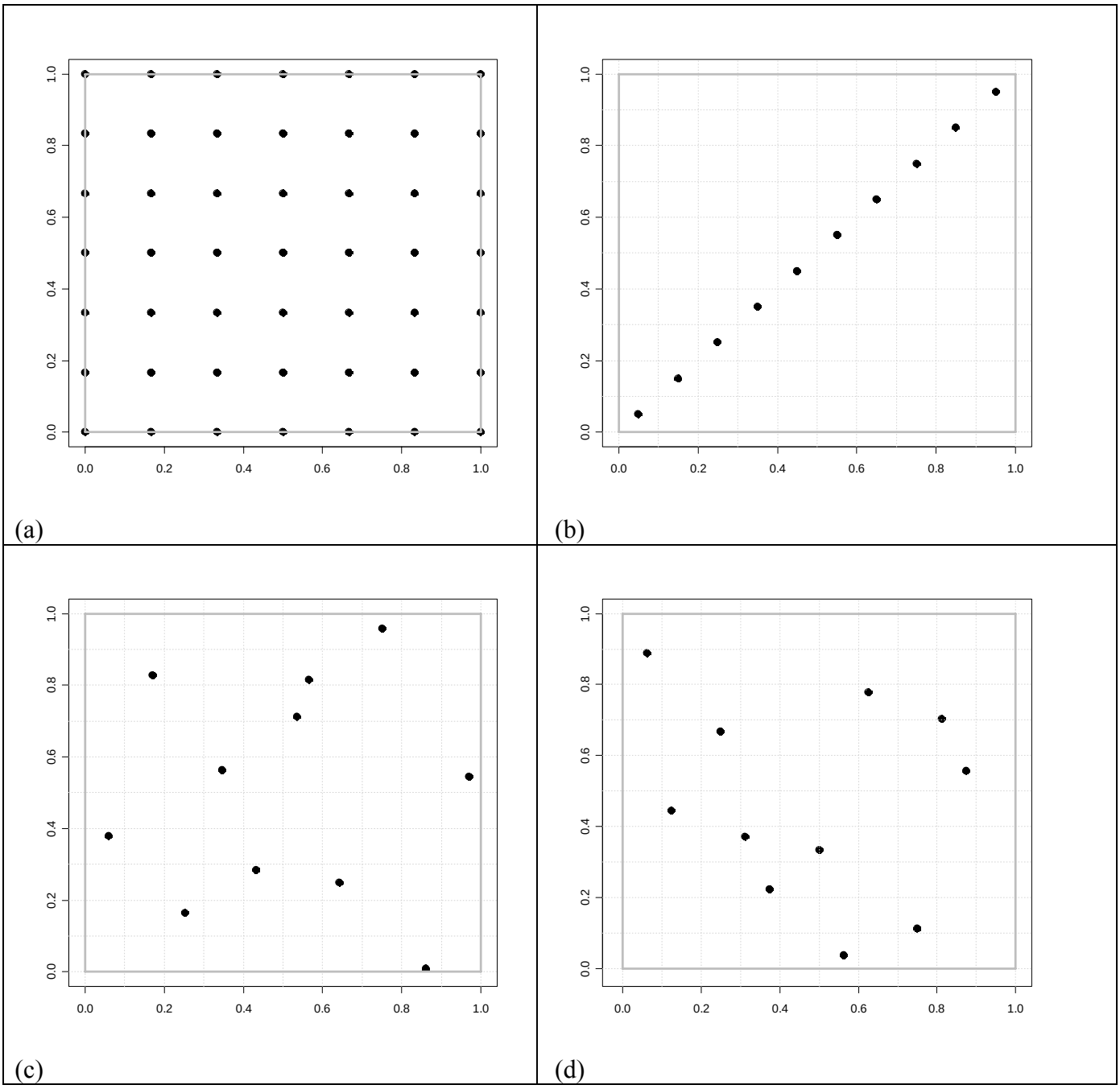
Numerical experiments are quite different from physical experiments in that repeated observations with the same set of inputs yield the same response (unless if the numerical code is changed). Uncertainty comes either from our lack of knowledge of the inputs or from the fact that the relationship between inputs and outputs is not perfectly known (or cannot be modelled exactly).
In this context, the selection of a design should take into account that replications (i.e. repeated observation at the same point) are not necessary and that the design should fit a large variety of models and provide information about all portions of the experimental design (i.e. no holes).
To do so the most useful designs are the ones satisfying the following two properties:
- to fill the space in order to be able to capture the nonlinearities of the numerical code (space filling designs – SFD)
- to conserve the filling property when projecting on sub-spaces, because often the numerical code depends only on few input variables (or on few linear combinations of the inputs).

The quality of the spatial repartition is measured either by deterministic criteria (minimax or maximin distances, see for instance [Johnson 90]) or by statistical criteria (such as discrepancy, see [Hickernell 98], [Niederreider 87]).

A non exhaustive list of designs of experiments is:

- Standard designs (see [Myers 02]) – even if they have good space filling properties, their main inconvenient is that they are time consuming for large number of inputs and they may lead to a loss of information in the context of numerical experiments
  - Full factorial designs (This is the simplest example of design, which consists in choosing the points on a regular grid with k levels for the d inputs. This leads to k^d simulations to be performed, which is inconceivable in the case of a large number of inputs, even for $k = 2$ or 3)
  - Fractional factorial design (see [Myers 02]
  - Doehlert design (see [Doehlert 70])
- Designs generated by latin hypercube sampling
  - Latin hypercubes, latin hypercubes satisfying some optimality criteria (such as maximin distance [Park 94] or correlation [Tang 98])
  - Orthogonal arrays [Owen 92]
  - Latin hypercubes based on orthogonal arrays [Tang 93]
- Designs based on low discrepancy sequences (see [Thiemard 00] for a unified construction of these sequences): Halton, Hammersley, Sobol, Faure.

Some examples of designs of experiments are given in Figure 39; they were generated by different R packages[2].



[2] R-packages can be downloaded at http://cran.r-project.org/

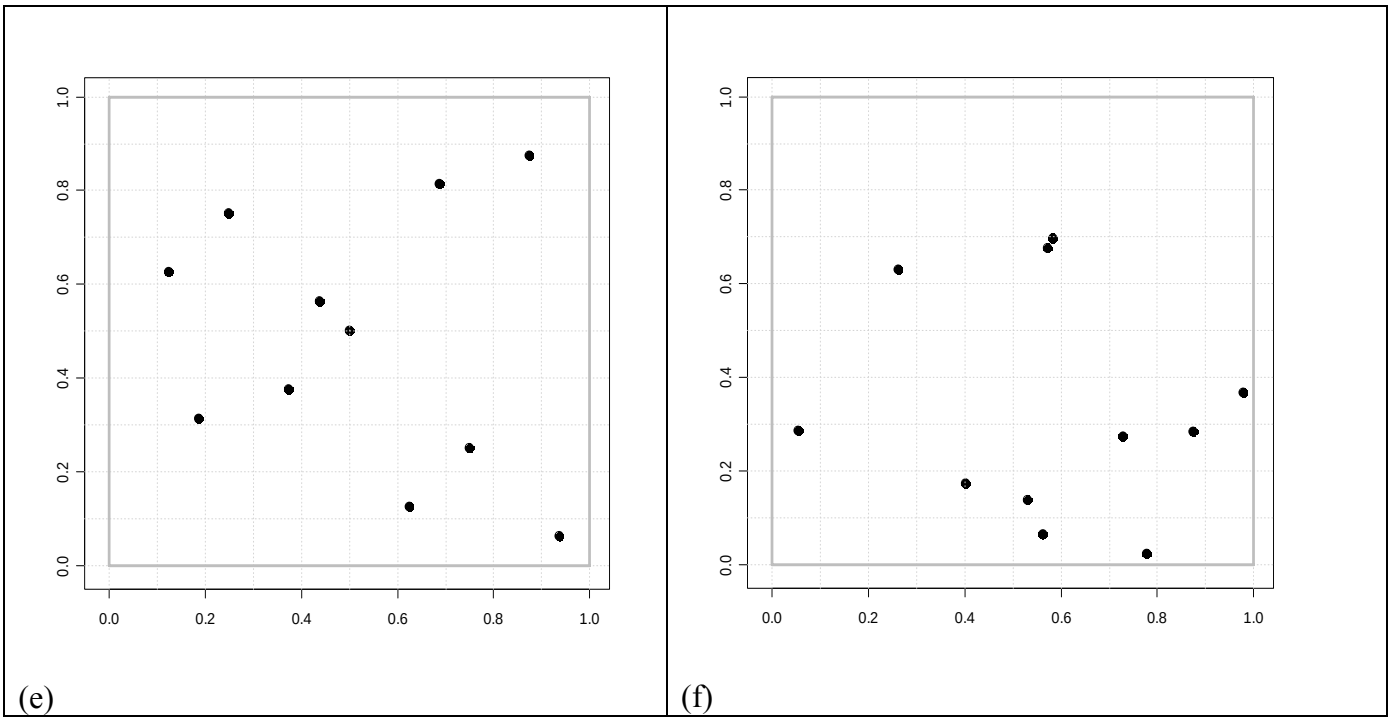


Figure 39: Some examples of designs of experiments in 2 dimensions : (a) full factorial design with 5 levels, i.e. 25 points, (b) latin hypercube with 10 levels (not space filling), (c) latin hypercube with maximin criteria (i.e. maximizing the minium distance between design points), (d) Halton sequence based design (low discrepancy) with 11 points, (e) Sobol sequence based design (low discrepancy) with 11 points, (f) Pseudo random design with 11 points.

5.5.2. The metamodels

We give here a very brief overview of some of the metamodels that can be used. As we previously mentioned the notation used here for a metamodel is

$$y(\boldsymbol{x}) = f(x_1, \ldots, x_d) + \varepsilon \ . \tag{5.22}$$

However, some more precise notation will be used in the following:
$\boldsymbol{x} = (x_1, \ldots, x_d)$ an experiments or a point of the design of experiments
$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$: a sample of n experiments; each point $\boldsymbol{x}_i$ having *d* coordinates.
$\boldsymbol{X}$: the matrix *n* x *d* of the design of experiments

Linear models (LM)

Linear models are written in the form

$$y(\boldsymbol{x}) = \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \ldots + \beta_M f_M(\boldsymbol{x}) + \varepsilon \tag{5.23}$$

where $\varepsilon \sim N(0, \sigma^2)$ is a white noise independent of x_i.
The model is linear because it is linear in its parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_M)$. The number of parameters to be estimated from the *n* experiments is equal to *M*+2: there are (*M*+1) β parameters (for the *trend*) and one σ^2 parameter (for the *dispersion*).
The functions $f_1, \ldots, f_M$ are predefined. They can be the input variables, some transforms of those variables (such as logarithms, powers, square roots…), functions of several variables. Their choice is

made before performing the estimation, leading to a more or less complex model. The simplest models are for instance

$$y(\boldsymbol{x}) = \beta_0 + \varepsilon \ , \tag{5.24}$$

and

$$y(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \varepsilon \, . \tag{5.25}$$

The estimators for the parameters β and σ^2 are computed by maximum likelihood:

$$\hat{\beta} = (F^T F)^{-1} F^T \boldsymbol{y} \tag{5.26}$$

and

$$\hat{\sigma}^2 = (\boldsymbol{y} - F\hat{\beta})^T (\boldsymbol{y} - F\hat{\beta}) / (n - (M+1)) \tag{5.27}$$

where *F* denotes the model's matrix[3]. Moreover their probabilistic distributions are known: $\hat{\beta}$ is a non-biased (i.e. it's expectation is β) Gaussian vector of variance $\sigma^2 (F^T F)^{-1}$ and $(n - (M+1))\,\hat{\sigma}^2 / \sigma^2$ is following a $\chi^2_{n-(M+1)}$ distribution.

Statistical tests about the model parameters are performed to measure the usefulness of the model.

The test for significance of regression is a test to determine whether a linear relationship between the response *y* and the regressors $f_1, \ldots, f_M$ exists .The appropriate hypotheses are that all the regressors are not contributing to the model (i.e. $\beta_1 = \beta_2 = \ldots = \beta_d = 0$) (H0) against the fact that there exists at least one regressor contributing significantly to the model (at least one of the β_j is different of 0) (H1). The test procedure is called analysis of variance (ANOVA) because it is based on the decomposition of the total variability of the response *y* and it uses the F-statistic (Fisher statistic).

Tests on individual regression coefficients are used to determine if the model is more effective with the inclusion of additional regressors or with the deletion of some variables already in the model. The appropriate hypotheses are that an individual regressor is not significant (β_j=0) (H0) against the fact that it is significant ($\beta_j \neq 0$) (H1). The test procedure uses the t-statistic (Student statistic).
For more details on linear models see [Myers 02].

After estimating the model's parameters, the model can be used for prediction purposes. Moreover, the variance of the prediction at a given point can also be computed analytically.

However there are situations in which a simpler model (with fewer terms) may be superior to the full model (i.e. the one with all the terms). **Variable selection** techniques have been developed to identify the best subset of regressors in a model. One procedure is to perform all the possible regressions, to evaluate them according to some criterion and to select the best regression model. Its main inconvenient is the fact that it is very time consuming. Hence some stepwise procedures have been developed for evaluating only a small number of subset regression models by adding or by deleting regressors one at a time. The stepwise procedures are either forward selection (regressors are added one at a time and the tests based on some F-statistic are performed), or backward elimination (starting from a model with all regressors, it removes the regressors one by one, based on the F-statistic), or a combination of those two methods working both ways. The stepwise procedures may also be based on AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) [Hastie 02].

---

[3] $F = \begin{bmatrix} 1 & f_1(\boldsymbol{x}_1) & \ldots & f_M(\boldsymbol{x}_1) \\ 1 & f_1(\boldsymbol{x}_2) & \ldots & f_M(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & f_1(\boldsymbol{x}_n) & \ldots & f_M(\boldsymbol{x}_n) \end{bmatrix}$

Additive (AM) and generalized additive (GAM) models

These models rely on the assumption of additivity:

$$y(\boldsymbol{x}) = \beta_0 + f_1(x_1) + \ldots + f_d(x_d) + \varepsilon \qquad (5.28)$$

where $\varepsilon \sim N(0, \sigma^2)$ is a white noise independent on x_i (some additional condition on f_i has to be considered in order to insure the uniqueness of the decomposition). The AM models have been introduced by Hastie and Tibshirani in [Hastie 90].
The form of the functions to be estimated is one of the parameters of the procedure, however polynomials or splines are most widely used.
The fitted model is formed by d one dimensional functions, describing the part of every predictor in the model. The main advantage of this type of decomposition is that it allows to fit non parametric models which are easy to interpret. Each component of the model represents an individual effect and shows how the expectation of the output y evolves as a function of one variable (predictor), all the other variables being fixed.
If the interactions are important, this type of model leads to wrong models. However, the model may be generalized (leading to GAM), by taking into account functions of several variables instead of functions of one single variable (such as functions of two variables as in [Hastie 90]).
In practice, the additive models are also a useful tool to visualise the variation of the output in the direction of each input. It allows obtaining some sort of a priori information which can be used to fit another type of model (such as the trend in a kriging model).
The stepwise procedures are also applicable and allow simplifying the model.

Kriging

Kriging was developed initially by geostatisticians in the 60's. It's use in the context of computer experiments started with the work of Sacks et al [Sacks 89a]. A detailed description of kriging may be found in [Santner 03]. Kriging is an interpolating method. The kriging model can be written in the form

$$y(\boldsymbol{x}) = f(\boldsymbol{x})\beta + Z(\boldsymbol{x}), \qquad (5.29)$$

where $f(\boldsymbol{x})\beta$ is the deterministic part (called the trend) and $Z(\boldsymbol{x})$ is the random part (a centered, stationary Gaussian process).
The trend takes into account the large scale variations of the output, while the random part takes into account its small scale variations.
The Gaussian process is entirely characterized by its covariance function. The covariance function determines the smoothness of the response surface and it depends on d+1 parameters (d being the number of inputs); its form has to be chosen a priori (out of a list of possible covariance functions, see [Santner 03]).

The parameters which have to be estimated from the data are the coefficients β of the trend and the d+1 parameters of the covariance function. The estimation is done by the maximum likelihood method.

As an example, the Gaussian covariance function is defined by:

$$C(h) = \sigma_Z^2 \exp\left(-\sum_{i=1}^{d} (h_i/\theta_i)^2\right) \quad \text{for any } h \in IR^d \qquad (5.30)$$

where σ_Z^2 and θ_i, i=1,…,d represent the variance and the scale parameters of the Gaussian process, have to be estimated from the experiments by maximum likelihood.

One of the main advantages of kriging is the fact the predictor is an interpolator whose smoothness depends on the data (via the covariance function) and yet not imposed a priori as in a regression method. The prediction variance may also be computed.

5.5.3. Model validation

When we are estimating a model (response surface), no matter what kind of model we are using, we need first to select a model and then to validate it. To select a model means to estimate the performances of different models in order to choose the best one. Once a model has been selected, validating it means to estimate its prediction error on independent data (i.e. data which have not been used to fit the model). Normally, on a set of data, one should use 2/3 of them to fit the model (these data are sometimes called calibration or training set) and keep 1/3 of them to validate it (these data are sometimes called validation set). If there is not enough data to do so, then one should use cross validation or bootstrap to validate (or select) the model (see [Hastie 02]). Other analytical criteria such as AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) [Hastie 02] can also be used.
One has also to be aware of the bias – variance tradeoff. This means that within a given family of models, a model with a low bias (i.e. a very complex model, depending on a large number of parameters) has a large variance and a model with low variance (i.e. a model depending on few parameters) has large bias. The best model (i.e. with the lowest Mean Squared Error) will have neither a very low bias nor a very low variance; the user will have to find the right tradeoff between biais and variance, meaning the right number of parameters to be used.
The more the complexity of the model increases, the more **the prediction error decreases on the training set**; unfortunately, this only applies to a certain degree, after a certain degree of complexity, **the prediction error increases on the validation set**. Just think at the extreme case of a polynomial of degree *n* passing through every point of the training set and providing an error equal to 0: this model is too sensitive to the sample (training set) itself. On the validation set, the error will be positive. This means that the predictive qualities of the model are bad. The typical description of this tradeoff is summarized in Figure 40. More details on this issue can be found in [Hastie 02].

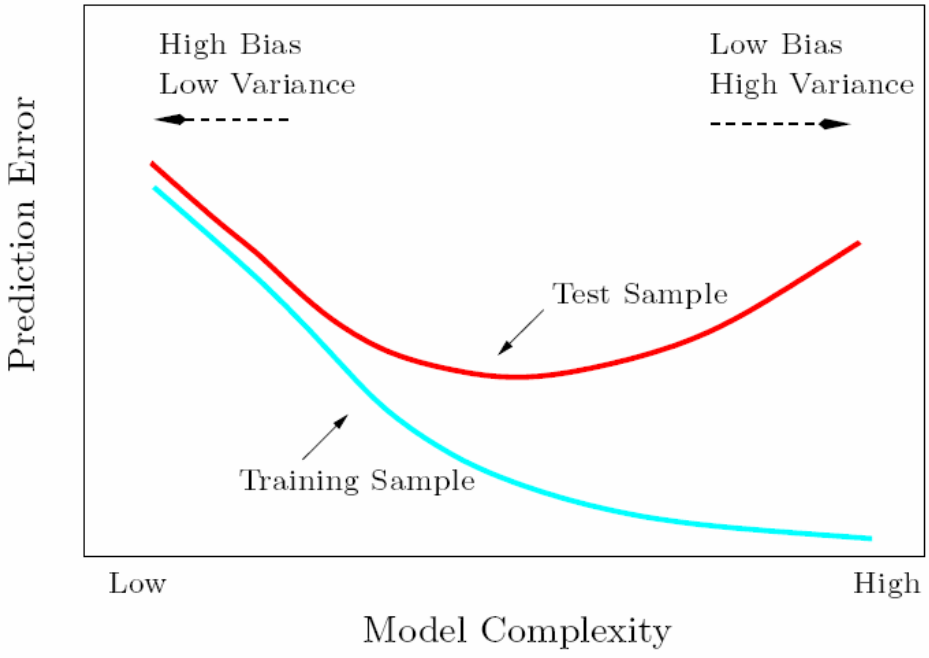


Figure 40: The tradeoff biais – variance (from [Hastie 02])

5.5.4. Example

In the following we present an example where the three types of metamodels from the previous section are applied. The example deals with the transport of one nuclide (^{129}I) in a nuclear waste repository and the results have been obtained in [Badea 07]. The computation is restricted to a 2D section of the disposal site, which has three different geological layers; the nuclear waste being disposed in the first one (the deepest one).
The original numerical code used is CAST3M (http://www-cast3m.cea.fr/) and the computation concerning the design of experiments and the metamodels have been performed using R (http://cran.r-project.org/).
The input of the numerical code consists of six environmental parameters:

- Kh1 : horizontal permeability of the first layer
- Kv1 : vertical permeability of the first layer
- K2 : permeability of the second layer
- K3 : permeability of the third layer
- poro : effective porosity
- de : effective diffusion coefficient

The output is the maximal release (concentration) of ^{129}I (denoted cI129) between 0 and 10^6 years at a predefined point located on the top of the third layer (and called exutory).

For each linear and GAM metamodel some statistical indicators such as the R^2 and R^2_{adj} coefficients have been computed (see for instance [Myers 02] for a detailed explanation of the significance of these coefficients). R^2, called the coefficient of multiple determination, represents the proportion of the total variation in the output explained by the model. One has to be aware of the fact that a large value of R^2 (close to 1), does not necessarily means that the model is good. Adding a new term to the model always increases the R^2, even if the added term is not statistically significant. It is hence possible for models with large values of R^2 to yield poor predictions of new observations. R^2_{adj} also takes into account the number of terms in the model; if unnecessary terms are added, the value of this coefficient will often decrease.
When the two R^2 and R^2_{adj} coefficients are very different, there is a good chance that nonsignificant terms have been included in the model.

Another computed quantity is the root mean square error (RMSE).

The R^2 and R^2_{adj} coefficients and the RMSE are not computed for interpolation models (in this case kriging models) for the set that have been used to estimate the model.

The design of experiments

A space filling design (latin hypercube with maximin distance) with 60 points have been used for this example. The distributions for the input parameters are known : log-normal for the permeabilities and the effective diffusion coefficient and uniform for the effective porosity. In Figure 41 we represent the design of experiments as a matrix of scatterplots.

Figure 41 : The design of experiments for the 6 inputs example

The data have been standardized using the sample mean and the sample standard deviation (i.e. for each variable the sample mean have been subtracted and the difference have been divided by the sample standard deviation).

The linear models

As we have indicated in the paragraph ***Linear models (LM)*** statistical tests for the significance of regression and on individual regression coefficients are performed for each linear model under study. Usually, the results of regression are presented as a table with 5 columns indicating respectively: the name of the variable (first column), the estimate of the corresponding β_j coefficient (second column), the standard error associated to the estimate of the coefficient (third column), the value of the t-statistic for H0 (i.e. β_j=0) of the test on individual regression coefficients (fourth column) and the corresponding *p*-value (fifth column). Sometimes (as it is the case here) an additional column containing qualitative information (such as "*" and ".") on the significance of the individual coefficients (going from "***" for the most significant to nothing for the least significant) appears in the table.

We present here for each of the two models the tables for the estimated coefficients classified by increasing p-value: the smaller the p-value, the more significant is the corresponding term in the model.

<u>*Model 1*</u>

This model contains the first degree terms and the interactions (i.e. the products between two different inputs). The estimated values of the coefficients β associated with each term are given in Table 4. The significance of "Intercept" is β_0. The symbol ":" means that a product between two terms is taken into account in the model.

Alternatively, the model may be written in the form:

$$
\begin{aligned}
cI129 = {} & -0.0802 - 0.0478\,Kh1 + 0.2294\,Kv1 + 0.0167\,K2 - 0.0287\,K3 - 0.1558\,poro - 0.066\,de \\
& - 0.119\,Kh1 \times Kv1 - 0.0417\,Kh1 \times K2 + 0.0325\,Kh1 \times K3 + 0.0132\,Kh1 \times poro + 0.0442\,Kh1 \times de \\
& + 0.0095\,Kv1 \times K2 + 0.0381\,Kv1 \times K3 - 0.3569\,Kv1 \times poro - 0.3189\,Kv1 \times de \\
& - 0.0456\,K2 \times K3 - 0.0192\,K2 \times poro + 0.028\,K2 \times de \\
& + 0.0286\,K3 \times poro - 0.0514\,K3 \times de \\
& - 0.0509\,poro \times de
\end{aligned}
\tag{5.31}
$$

	Estimate	Std.Error	t-value	p-value (Pr(>|t|))	
Kv1:poro	-0.3569	0.0447	-7.984	1.20E-09	***
Kv1:de	-0.3189	0.0542	-5.886	8.15E-07	***
poro	-0.1558	0.0346	-4.503	6.18E-05	***
Kv1	0.2294	0.0519	4.42	7.96E-05	***
(Intercept)	-0.0802	0.0334	-2.4	0.0214	*
Kh1:Kv1	-0.1190	0.0615	-1.936	0.0604	.
de	-0.0660	0.0389	-1.696	0.098	.
K3:de	-0.0514	0.0354	-1.452	0.1547	
Kh1	-0.0478	0.0387	-1.234	0.2249	
K2:K3	-0.0456	0.0455	-1.003	0.3223	
poro:de	-0.0509	0.0517	-0.985	0.331	
Kh1:de	0.0442	0.0511	0.865	0.3923	
K3	-0.0287	0.0339	-0.846	0.403	
Kh1:K3	0.0325	0.0386	0.841	0.4059	
K3:poro	0.0286	0.0356	0.802	0.4277	
Kh1:K2	-0.0417	0.0562	-0.742	0.4625	
Kv1:K3	0.0381	0.0574	0.664	0.5109	
K2:de	0.0280	0.0437	0.641	0.5257	
K2:poro	-0.0192	0.0345	-0.555	0.5818	
K2	0.0167	0.0397	0.421	0.6759	
Kh1:poro	0.0132	0.0517	0.256	0.7992	
Kv1:K2	0.0095	0.0816	0.117	0.9078	

Table 4 : Coefficients of the linear model with first degree polynomial and interactions; classified by increasing p-value

The statistical indicators for this model are:

$R^2 = 0.9675$
$R^2_{adj} = 0.9495$
RMSE = 0.1788

One should be aware that such a model (5.31) has far too many terms (22), most of which are not statistically significant and should not be used for prediction. Instead, a stepwise procedure should be used to eliminate the terms that are not significant.

Usually a detailed study of the residuals (difference between real and predicted values) should also be performed. Hereafter we only present a qqplot picture of the residuals.

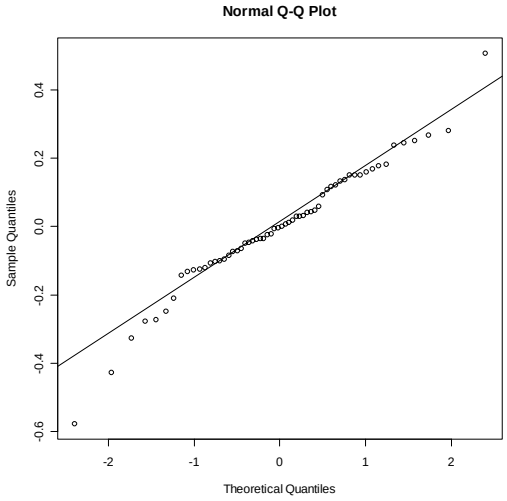


Figure 42 : qqplot for the residuals in Model 1

Model 2

Starting from Model 1, a stepwise procedure (backward elimination) has been applied to obtain a new model and only 10 term have been kept out of the initial 22 terms.

The model may be written in the form:

$$cI129 = -0.07297 - 0.01938Kh1 + 0.24954Kv1 - 0.0298K3 - 0.14555poro - 0.07168de - 0.08453Kh1 \times Kv1 - 0.33933Kv1 \times poro - 0.33212Kv1 \times de - 0.04793K3 \times de \quad (5.32)$$

	Estimate	Std.Error	t-value	Pr(>|t|)	
Kv1:poro	-0.33933	0.03471	-9.776	3.42E-13	***
Kv1:de	-0.33212	0.04234	-7.845	2.89E-10	***
Kv1	0.24954	0.04353	5.733	5.67E-07	***
poro	-0.14555	0.02807	-5.184	3.92E-06	***
(Intercept)	-0.07297	0.02799	-2.607	0.012	*
de	-0.07168	0.03274	-2.189	0.0333	*
Kh1:Kv1	-0.08453	0.04388	-1.926	0.0598	.
K3:de	-0.04793	0.02525	-1.898	0.0634	.
K3	-0.0298	0.02774	-1.074	0.2879	
Kh1	-0.01938	0.02794	-0.694	0.4912	

Table 5 : Coefficients of the linear model obtained from the previous linear model using a stepwise procedure

The statistical indicators for this model are:

$R^2 = 0.9635$
$R^2_{adj} = 0.9569$
RMSE = 0.1894

This model has only 10 terms and has larger R^2_{adj} than Model 1. For these reasons Model 2 should be preferred to Model 1.

The GAM models

Model 3

For this example, the f_i functions are splines. The table below only shows the degree of significance of each spline.

	Pr(F)	
s(Kv1)	2.20E-16	***
s(poro)	1.46E-05	***
s(K3)	0.1309	
s(Kh1)	0.3418	
s(K2)	0.6553	
s(de)	0.7049	

Table 6 : Degree of significance of each spline, ordered in increasing order of the p-value

The statistical indicators for this model are:

$R^2 = 0.9974$
$R^2_{adj} = 0.9965$
RMSE = 0.051

One should be aware of the fact that the number of degrees of freedom of this model is 24 (while for Models 1 and 2 these numbers were 21 and 9 respectively).

Each component of the model (representing an individual effect) shows how the expectation of the output evolves as a function of one variable. This can be seen in Figure 43.

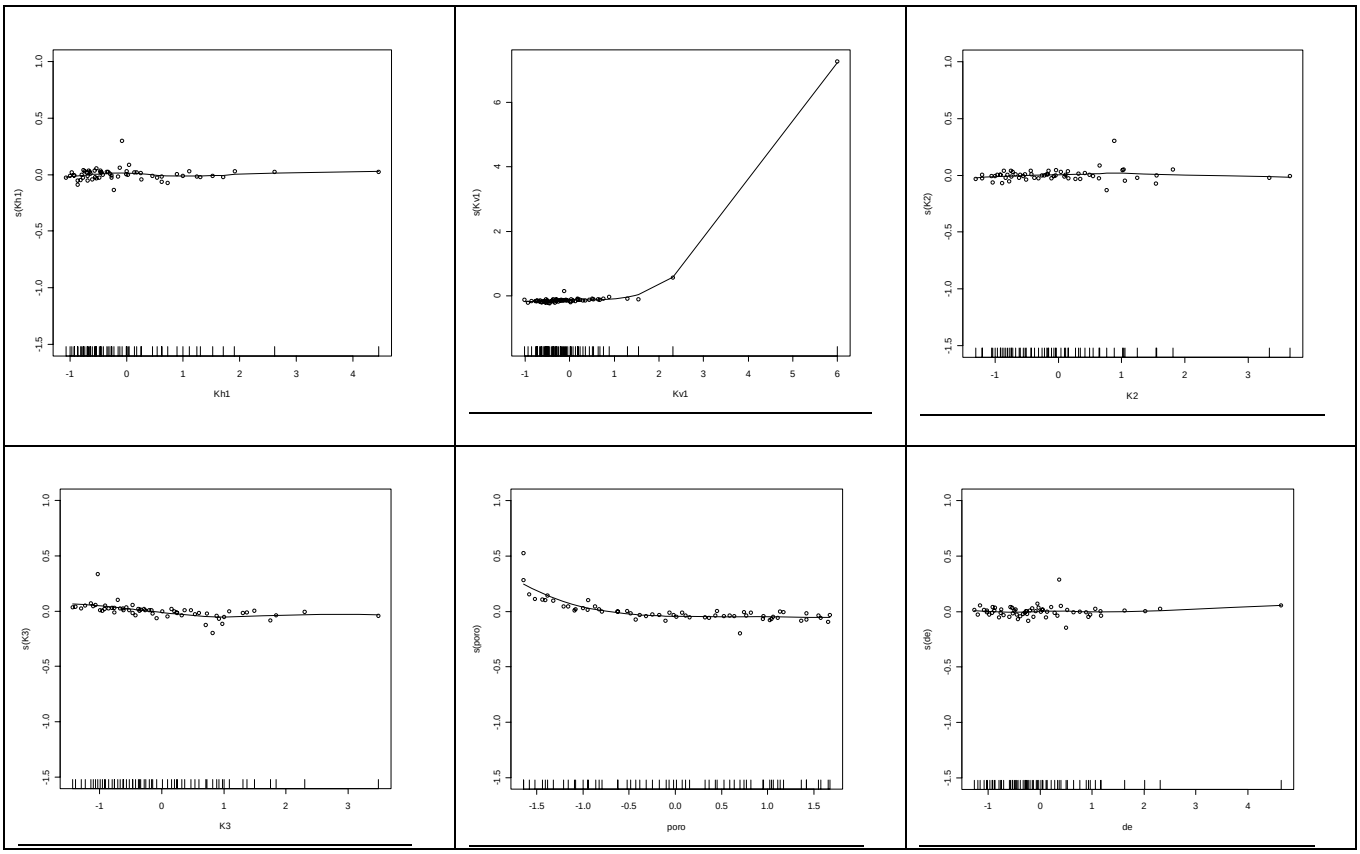


Figure 43 : The output as a function of each input: Kh1, Kv1, K2, K3, poro and de (from left to right and from top to bottom)

One may remark that most of the effects are linear, except for the parameters Kv1 and poro, which were also the most significant parameters of this model.

Model 4

By applying a stepwise procedure, a new model with only three functions is obtained.

	Pr(F)	
(Intercept)		
s(Kv1)	2.20E-16	***
s(poro)	8.59E-07	***
s(K3)	0.07049	.

Table 7: Degree of significance of each spline for the stepwise model

The statistical indicators for this model are:

$R^2 = 0.9969$
$R^2_{adj} = 0.9961$
RMSE = 0.0555
The number of degrees of freedom of this model is 12.

The kriging models

Model 5

An ordinary kriging model has been estimated, which means that the deterministic part consists only of a constant β which has to be estimated. The covariance function associated with the random part is Gaussian (as defined in (5.30)); its d+1=7 parameters have also to be estimated.

	θ	σ_Z	β
		1.8	1.2
Kh1	40		
Kv1	3.59		
K2	25.75		
K3	11.42		
poro	2.11		
de	40		

Table 8: Estimated parameters for the ordinary kriging model

The number of degrees of freedom of this model is 8.
The main remark here is that some of the scale parameters of the covariance function (i.e. the ones denoted by θ) are quite large, taking into account that the data have been standardized. The meaning is that the response surface is not varying (or is slightly varying) in the corresponding direction, the effect of the corresponding parameter being included in the deterministic part of the model. In this example only the parameters Kv1 and poro (the ones with low scale parameters) could be considered responsible for the variations in the random part. It should be emphasised that those parameters are the ones that appear as important in the other models too.

As previously mentioned, the statistical indicators R^2, R^2_{adj} and RMSE are meaningless for a kriging model. A leave-one-out cross validation may be performed; some results are shown in Figure 44, where the fitted values (i.e. predicted values) are plotted against the real ones. It can be seen that

nearly all the points are aligned on the line $y = x$ with one exception. This exception is due to the point number 44 of the design. The user should decide whether this point is an outlier or not and whether it should be kept in the design or not. Of course, all the models can be re estimated using only the remaining 59 points of the design.

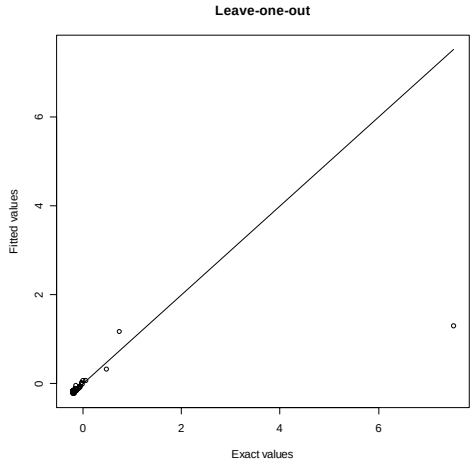


Figure 44 : Predicted against exact values for a kriging model

The comparison

A design with 737 points (on a regular grid) has been used as a validation set.

	Model 1	Model 2	Model 3	Model 4	Model 5
RMSE	0.2495	0.2452	0.0631	0.0631	0.2614
df	21	9	24	12	8

Table 9: Comparison of the 5 models in terms of RMSE on the validation test and in terms of degrees of freedom

Some other statistical indicators may be computed (see [Myers 02]) in order to decide which model to select. We only present here the RMSE and the degrees of freedom of each model. One should choose a model with not too many degrees of freedom, with reasonable RMSE. In this case either Model 2 or Model 5 should be preferred, also because we have seen that interactions between parameters cannot be neglected (and the GAM models do not take it into account).

6. Uncertainty of the output

The outputs of numerical codes may be either scalars or functions of other variables (most often time dependent functions).

6.1. Case of a scalar output

In this case the standard procedures for descriptive statistics apply, and we compute the standard indicators: mean, variance, median, different quantiles, as well as the PDF estimation and the empirical (complementary) cumulative distribution function for the output of interest. In section 6.2.2 we will address, via an example most of those standard indicators.

6.2. Case of a functional output

The case of a functional output will be handled by means of an example.
Let us consider the example we have already started to present throughout this report. It is a study of the release of radionuclides from wastes of an ILW disposal cell (represented in Figure 45) embedded in a porous material for a generic French clay site, in the case of a simplified 2D representation of the disposal cell. This study has been performed in [Prváková 08]. There are 24 uncertain input parameters. The output consists of 18 dynamic variables (time dependent functions) and 36 non-dynamic variables (scalar variables).

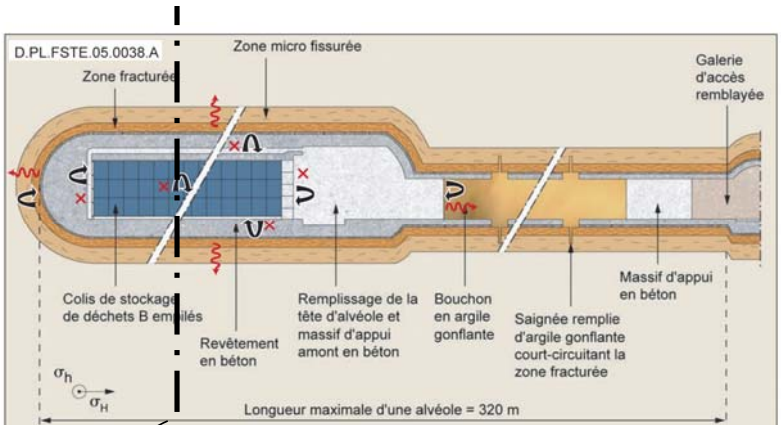


Figure 45: Disposal cell scheme

The 18 dynamic variables are the molar flows (for each radionuclide (^{129}I, ^{94}Nb, ^{79}Se) released through the external surface of each of the 6 engineered and geological barriers[4] considered in the repository design) as time functions (time being considered from 0 to 10^6 years).

The 36 non-dynamic variables are the peaks of the previous molar flows (18 variables) and the time to the peak (18 variables).

The number of simulations by Monte Carlo is 1000; for each one, the results for 209 time steps have been stored. The numerical code used to perform the simulations is Goldsim (http://www.goldsim.com/).

More non-dynamic variables may be considered, as for instance the molar flow for one radio nuclide for a given time.

6.2.1. Example of study of dynamic variables

For each dynamic output we extend the standard procedure for the scalar output to the functional output by considering the evolution in time of the statistical indicators (mean, variance, median, different quantiles), see for instance Figure 46.

[4] The 6 barriers are: waste package, disposal cell, fractured zone, micro fissured zone, undisturbed argillites (15m from top and bottom) and undisturbed argillites.

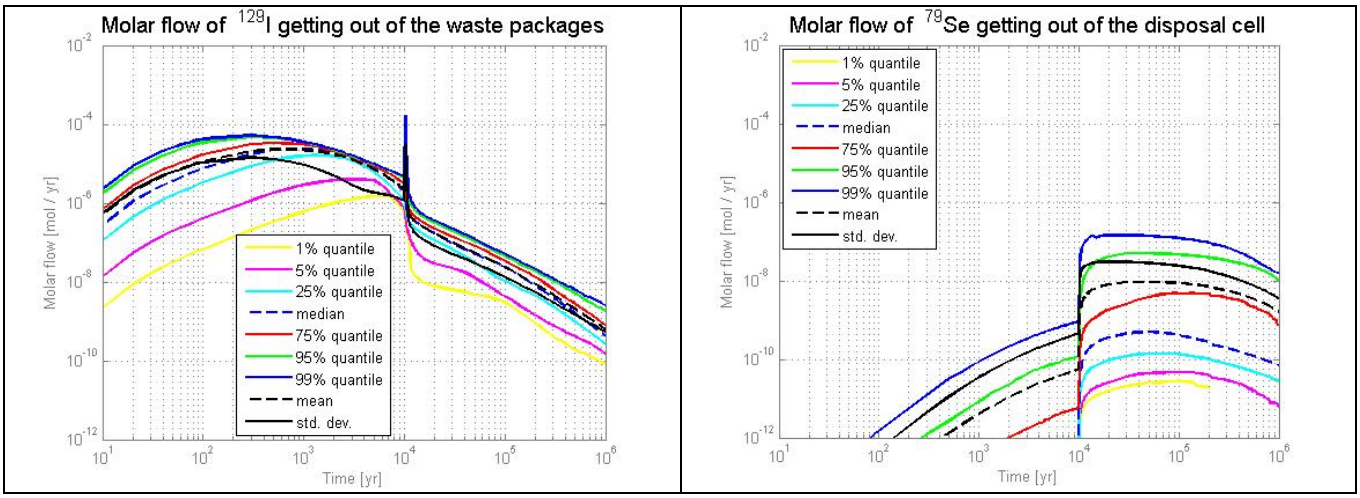


Figure 46: Evolution in time of the main statistical indicators for different radionuclides coming out of different surfaces

In this example, in addition to the simple knowledge of the values of the statistical indicators represented in Figure 46 at different moments, one can see a sudden increase in the molar flow, producing peaks at 10^4 years. This is due to the change, at that precise moment, of the distribution of one of the input parameters. In other studies, other features may appear on the evolution in time of different statistical indicators curves.

At each moment, for every statistical indicator a confidence interval may be computed (with a given confidence level). This yields an evolution in time of the confidence interval for the specified indicator, as can be seen in Figure 47. One should be aware of the fact that the confidence interval is computed for each fixed instant (with the significance given in 4.1.2).

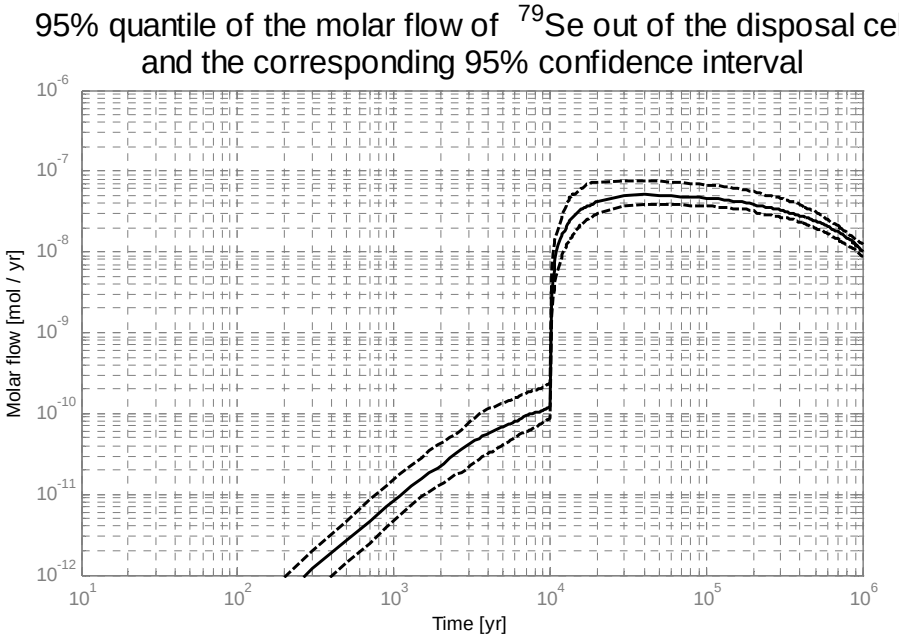


Figure 47 : Evolution in time of the 95% confidence interval for the 95% quantile of the molar flow of ^{79}Se getting out of the disposal cell

6.2.2. Example of study of non-dynamic variables

Non-dynamic variables (fixed instant)

For each dynamic output, we also consider their values at some predefined times (in this case we considered $5x10^3$ years, 10^4 years, $5x10^4$ years, 10^5 years, $5x10^5$ years, 10^6 years). Then the scalar analysis mentioned in paragraph 6.1 may be applied to each one of these times. We present in Table 10 the statistical indicators for the molar flow of ^{129}I at 10^5 years and in Figure 48 the ECDFs for the log of the molar flow of ^{129}I at 10^6 years coming out of each of the 6 barriers (left) and similarly the PDFs for the log of the molar flow of ^{79}Se at 10^5 years (right).

	min	1% quantile	5% quantile	25% quantile	median	75% quantile	95% quantile	99% quantile	max	mean	std.dev	skewness	kurtosis
waste package	**2.E-09**	**3.E-09**	**4.E-09**	**1.E-08**	**2.E-08**	**3.E-08**	**4.E-08**	**5.E-08**	**5.E-08**	**2.E-08**	**1.E-08**	**0.13**	**1.93**
disposal cell	9.E-09	1.E-08	2.E-08	6.E-08	1.E-07	2.E-07	2.E-07	3.E-07	3.E-07	1.E-07	7.E-08	0.26	1.94
fractured zone	2.E-08	4.E-08	7.E-08	1.E-07	2.E-07	2.E-07	3.E-07	3.E-07	4.E-07	2.E-07	7.E-08	0.21	2.47
micro fissured zone	3.E-11	7.E-09	3.E-08	2.E-07	3.E-07	3.E-07	4.E-07	4.E-07	4.E-07	2.E-07	1.E-07	-0.69	2.60
undisturbed argilites (15 m)	2.E-14	2.E-11	6.E-10	3.E-08	2.E-07	4.E-07	4.E-07	4.E-07	4.E-07	2.E-07	2.E-07	-0.08	1.37
undisturbed argilites	0.E+00	0.E+00	0.E+00	4.E-23	8.E-15	2.E-10	3.E-08	9.E-08	2.E-07	5.E-09	2.E-08	5.68	42.52

Table 10: Uncertainty analysis (numerical indicators) for the molar flow of ^{129}I at 100000 years

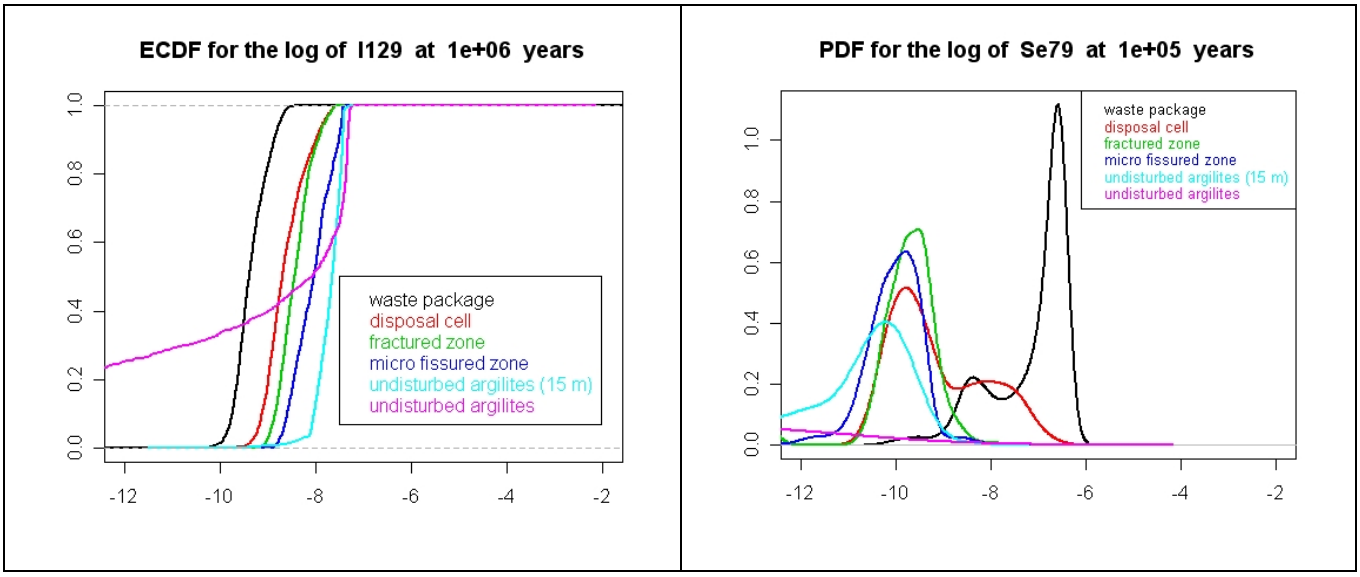


Figure 48 : Uncertainty analysis (graphics) : ECDF for the log of the molar flow of 129I at 1000000 years (left) and PDF for the log of the molar flow of 79Se at 100000 years (right).

We may also retrieve directly from the curves in Figure 46, the values of some statistical indicators at some fixed moment; each one of those values will characterize a non-dynamic variable. For instance,

at 10^5 years, the molar flow of ^{129}I getting out of the waste packages has a mean value and a median which are the same (around $2x10^{-8}$mol/yr) and a standard deviation of 10^{-8}mol/yr (Figure 46, left). The same quantitative information can be obtained directly from the second line of Table 10. Some additional information concerning the shape characteristics for this variable may be retrieved from this table: the skewness coefficient 0.13 tells us that that the distribution is rather symmetric, while the kurtosis equal to 1.93 leads to the conclusion that the distribution is quite flat.

On the contrary, at 10^5 years, the molar flow of ^{79}Se getting out of the disposal cell has a mean value ($8.6x10^{-9}$mol/yr), quite different from the median ($4x10^{-10}$mol/yr); the standard deviation is $2.4x10^{-8}$ mol/yr (see Figure 46, right). The statistical indicators for this output are given in Table 11.

min	1% quantile	5% quantile	25% quantile	median	75% quantile	95% quantile	99% quantile	max	mean	std.dev	skewness	kurtosis
1.76E-11	2.82E-11	4.72E-11	1.32E-10	4.08E-10	4.71E-09	4.63E-08	1.20E-07	2.86E-07	8.64E-09	2.38E-08	5.48	43.47

Table 11 : Statistical indicators for the molar flow of ^{79}Se getting out of the disposal cell at 10^5 years

From this table we can also derive the range, 2.85655×10^{-7} mol/yr, and the interquartile interval, equal to 4.582×10^{-9} mol/yr.
It should be observed that the skewness coefficient of 5.48 gives a strong indication of non-symmetric distribution, whereas the very large kurtosis (equal to 43.47) tells us that there is a strong proportion of the distribution located at the tails.

The ECDF and the PDF curves for the considered outputs are plotted in Figure 49.

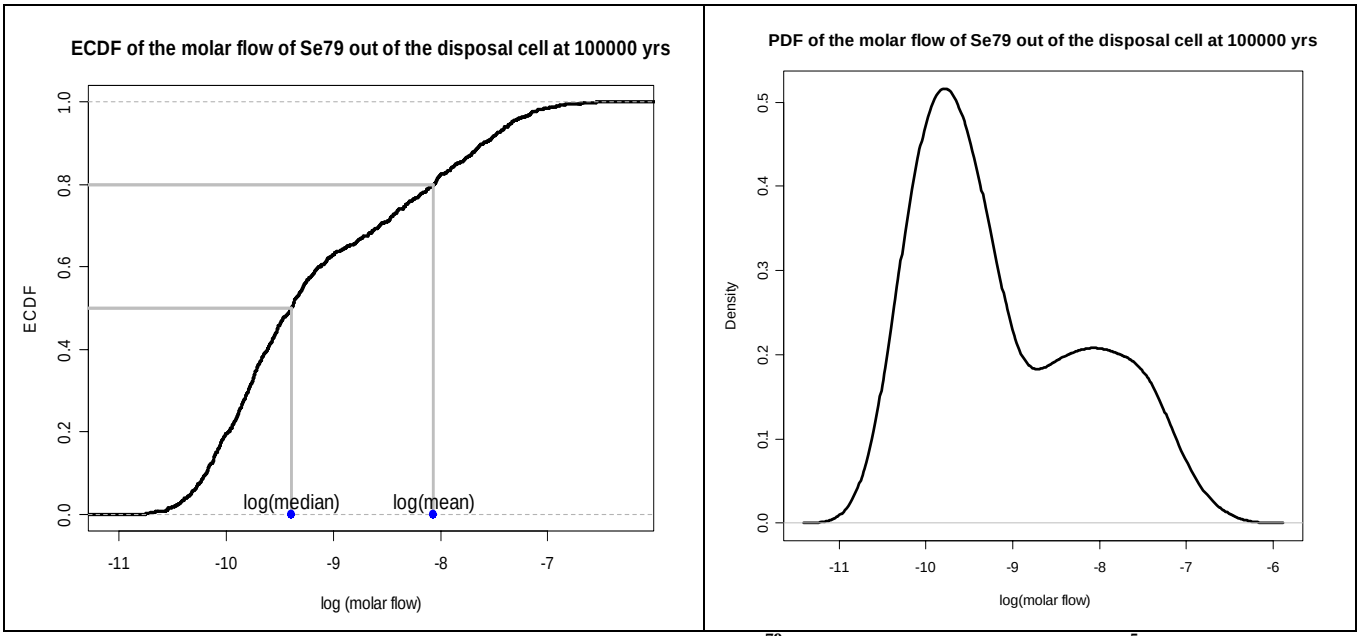


Figure 49: ECDF and PDF curves for the log of the molar flow of ^{79}Se out of the disposal cell at 10^5 years

From the s-shaped qqplot of log of the molar flow of ^{79}Se out of the disposal cell at 10^5 years, we can conclude, that its distribution is bi-modal (Figure 50).

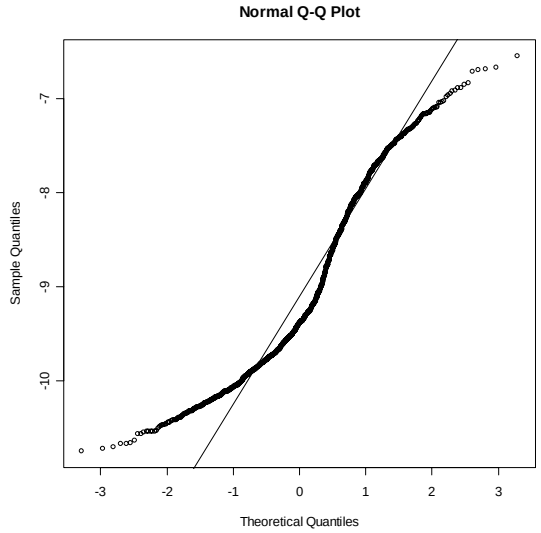


Figure 50: qqplot of log of the molar flow of ^{79}Se out of the disposal cell at 10^5 years

Non-dynamic variables (peak or time of occurrence of the peak)

For each dynamic output, we may also consider the peak (maximum) and the time of occurrence of the peak, which are, in the safety context the most important scalar outputs to be considered.

Examples of statistical indicators for the peak are given in Table 12 and Figure 51.
Examples of statistical indicators for the time of occurrence of the peaks are given in Table 13 and Figure 52.

Some comments should be done concerning the way to obtain the peaks. For each simulation with the numerical code results at 209 predefined time steps have been stored. The peak, for each simulation, has been chosen among the 209 values, without any interpolation, inducing hence a small biais. As for the time to the peak, only 209 values were possible, out of which much fewer were really taken. For this reason the time to the peak has the aspect of a discrete variable (see the stair-like shape of the ECDF in Figure 52, left, and the multi-modal PDF in Figure 52, right).

The main conclusions concerning these outputs are the following:

- for the distributions of peaks for ^{79}Se
  - the distributions shift towards smaller values as moving to outer layers (i.e. from the waste package towards the undisturbed argillite)
  - the median is smaller than the mean for all the layers
  - the standard deviation is also decreasing as moving to outer layers
  - the ratio between the range and the standard deviation is however increasing as moving to outer layers (from 11 in the waste package layer to 20 in the undisturbed argillite layer)
  - the distributions are not symmetric (large skewness coefficient) and have long tails (large kurtosis)
- for the distributions of time to the peaks
  - the distributions shift to larger values when moving to outer layers
  - the median is smaller than the mean for nearly all the layers
  - the standard deviation increases when moving to outer layers
  - for ^{94}Nb we can see from Figure 52 that in the undisturbed argillite layer the true peak of the molar flow has not been reached before the simulations stopped (there is no curve for this

layer); the same information can be obtained by checking the numerical values of the statistical indicators (the table is not presented in this report).

Of course, all the previous computed indicators (range, interquartile interval …) can also be calculated here.

	min	1% quantile	5% quantile	25% quantile	median	75% quantile	95% quantile	99% quantile	max	mean	std.dev	skewness	kurtosis
waste package	7.E-07	7.E-07	9.E-07	1.E-06	2.E-06	5.E-06	2.E-05	4.E-05	1.E-04	5.E-06	9.E-06	5.40	43.82
disposal cell	2.E-11	3.E-11	5.E-11	1.E-10	5.E-10	5.E-09	5.E-08	1.E-07	5.E-07	1.E-08	3.E-08	7.82	88.13
fractured zone	3.E-14	6.E-14	1.E-13	6.E-11	2.E-10	5.E-10	2.E-09	9.E-09	1.E-07	7.E-10	5.E-09	22.56	600.07
micro fissured zone	6.E-15	1.E-14	4.E-14	4.E-11	1.E-10	2.E-10	5.E-10	2.E-09	3.E-08	2.E-10	9.E-10	22.32	596.90
undisturbed argilites (15 m)	2.E-16	5.E-15	2.E-14	2.E-11	5.E-11	1.E-10	3.E-10	1.E-09	1.E-08	1.E-10	4.E-10	20.16	508.49
undisturbed argilites	0.E+00	2.E-29	1.E-24	4.E-17	3.E-13	1.E-11	7.E-11	1.E-10	1.E-09	2.E-11	5.E-11	11.97	195.01

Table 12 : Uncertainty analysis (numerical indicators) for the peak of the molar flow of ^{79}Se

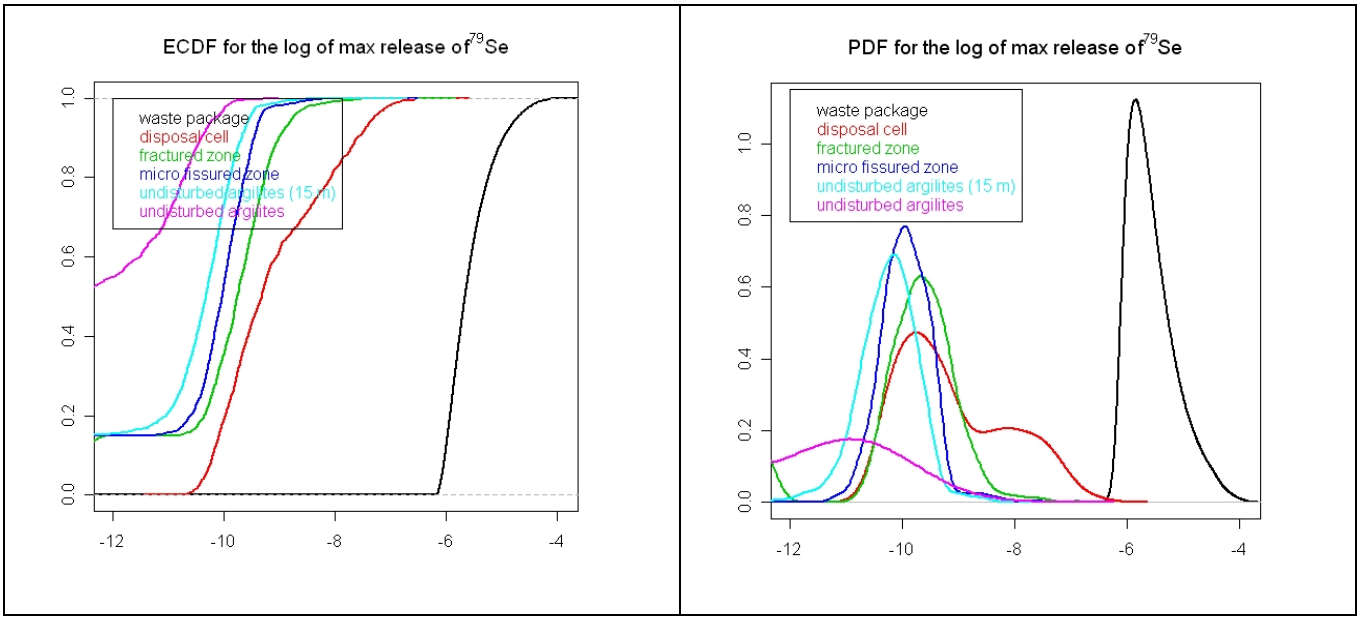


Figure 51: Uncertainty analysis (graphics) : ECDF (left) and PDF (right) for the log of the peak (max) of molar flow of ^{79}Se

	min	1% quantile	5% quantile	25% quantile	median	75% quantile	95% quantile	99% quantile	max	mean	std. dev	skewness	kurtosis
waste package	300	300	300	700	900	1500	7540	10200	10200	1659	2236	2.97	11.14
disposal cell	900	900	1100	1500	1900	2900	10200	10200	10200	2974	2617	2.14	6.22
fractured zone	1700	2100	2500	3300	4500	6200	10800	11402	13400	5525	2868	0.96	2.48
micro fissured zone	5800	6998	8790	13600	19800	70000	150000	230100	930000	49821	60448	4.36	49.22
undisturbed argilites (15 m)	13200	17400	20000	40000	95000	220000	800500	1E+06	1000000	191915	238244	2.05	6.59
undisturbed argilites	10	210000	280000	650000	1E+06	1E+06	1E+06	1E+06	1000000	826350	261962	-1.17	2.84

Table 13: Uncertainty analysis (numerical indicators) for the time to the peak of the molar flow of ^{129}I

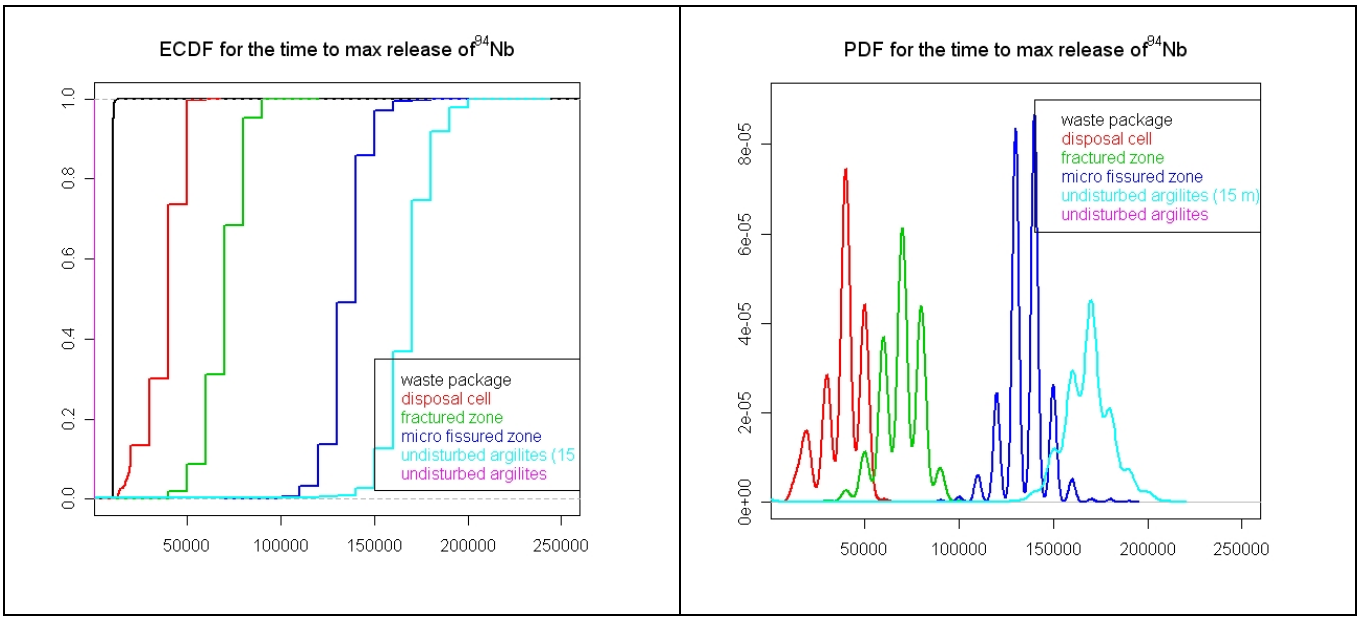


Figure 52: Uncertainty analysis (graphics) : ECDF (left) and PDF (right) for the time to the peak (max) of molar flow of ^{94}Nb.

6.2.3. Conclusion

To summarize, in the case of functional outputs for a PA code, one should analyze:

- concentrations of different radionuclides at a fixed point as functions of time
- molar flows of different radionuclides coming out of different specified surfaces (disposal cell for instance) as functions of time.

For each of these outputs the standard procedure for the scalar output can be extended to the functional output by building the evolution in time of the statistical indicators (mean, variance, median, different quantiles).

For each of these outputs, their values at some predefined moments (such as 10^4 years, 10^5 years, 10^6 years or others) should be considered. Then the scalar analysis may be applied to each one of these values.
The scalar analysis may be applied to the peak (maximum) of each time dependent output and also to the time of occurrence of the peaks.

7. Conclusion

Dealing with uncertainties for Performance Assessment of a nuclear High Level Waste repository (HLW) involves

- Assessment of the input uncertainty (which may be epistemic uncertainty due to the lack of knowledge and aleatory uncertainty) and assessment of the uncertainties due to the modeling of the whole system (which is rather complex)
- Uncertainty propagation using Monte Carlo method (which consists in repeating *n* times the two following steps : first generating a set of realizations of the uncertain inputs and secondly, for this set of realizations compute and store the model outputs)
- For each output (such as the maximum releases of radionuclides), assessment of its uncertainty from the *n* values of the model output.

The input uncertainty is expressed through a probability distribution characterized by its shape (uniform, normal, log-normal, exponential,...) and its parameters. The process of appropriately characterizing the uncertainty in the inputs may be time-consuming, especially when dealing with models involving tens (or even thousands) of inputs. In these situations it is helpful to perform a first uncertainty and sensitivity analysis with coarse uncertainty characterizations in order to identify the most important input variables and then concentrate on refining their distributions.
Correctly characterizing the uncertainty of the inputs is essential for any meaningful uncertainty analysis.

The accuracy of Monte Carlo method increases with the number of runs *n* (it is inversely proportional to the square root of *n*). It is therefore an expensive method and alternative methods (called variance reduction methods), such as Latin Hypercube sampling, may be used instead of MC.

In this document we summarised the main techniques available in scientific literature to perform uncertainty analysis. Some examples are also given.

8. Annex 1

8.1. Properties of the empirical estimator

The estimator of the α- quantile $\hat{Y}_{\alpha,n} = \inf\{y, \hat{F}_{EE}(y) > \alpha\} = Y_{(\lceil \alpha n \rceil)}$ is biased with the following first two moments

$\mathrm{E}(\hat{Y}_{\alpha,n}) = y_\alpha - \frac{\alpha(1-\alpha)p'(y_\alpha)}{2(n+2)p^3(y_\alpha)} + O\left(\frac{1}{n^2}\right)$, $\mathrm{Var}(\hat{Y}_{\alpha,n}) = \frac{\alpha(1-\alpha)}{(n+2)p^2(y_\alpha)} + O\left(\frac{1}{n^2}\right)$ and asymptotically normal

$\sqrt{n}(\hat{Y}_{\alpha,n} - y_\alpha) \xrightarrow{n\to\infty} N\left(0, \frac{\alpha(1-\alpha)}{p^2(y_\alpha)}\right)$. The variance of this estimator is large, and it increases for extreme quantiles, for which the value of the PDF f(y_a) is small. Moreover, from the asymptotical law, one can see that $P(\hat{Y}_{\alpha,n} \geq y_\alpha) \approx 0.5$.

8.2. Properties of the Wilks estimator

The Wilks estimator is the order statistic $Y_{(n-r+1)}$.

The following proposition helps establishing the Wilks formula, which connects *n* (the size of the sample) and r, α, β from $P(Y_{(n-r+1)} > y_\alpha) \geq \beta$.

Proposition :

The number of times n iid rv (Y_1, …, Y_n) exceed a certain threshold y follows a Binomial distribution B(n,F(y)), where F is the CDF of the rv Y_i.

This proposition is used as follows: let $(Y_{(1)},...,Y_{(n)})$ be the ordered sample, $Y_{(1)} \leq ... \leq Y_{(n)}$. The probability of the event $\{j$ of the Y_i are $> y\}$, $\forall j$ is computed using the binomial distribution :

$P(j \text{ of the } Y_i \text{ are} > y) = \binom{n}{j}(1-F(y))^j F(y)^{n-j}$.

For $y = y_\alpha$, we get $P(j \text{ of the } Y_i \text{ are} > y_\alpha) = \binom{n}{j}(1-\alpha)^j \alpha^{n-j}$. On the other hand, the event $\{Y_{(n-r+1)} > y_\alpha\}$ occurs if and only if at least *r* of the Y_i are > y_α. This leads to

$$P(Y_{(n-r+1)} > y_\alpha) = \sum_{j=r}^{n} P(j \text{ of the } Y_i \text{ are} > y_\alpha) = \sum_{j=r}^{n} \binom{n}{j}(1-\alpha)^j \alpha^{n-j}$$

$$= \sum_{j=0}^{n-r} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} = 1 - \sum_{j=n-r+1}^{n} \binom{n}{j} \alpha^j (1-\alpha)^{n-j}$$

As we want $Y_{(n-r+1)}$ to be sure to the level β, i.e. $P(Y_{(n-r+1)} > y_\alpha) \geq \beta$, we obtain the « Wilks formula »

$$\beta = 1 - \sum_{n-r+1}^{n} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} = \sum_{j=0}^{n-r} \binom{n}{j} \alpha^j (1-\alpha)^{n-j}$$

It is then possible to compute, for α, β fixed (for instance 95%, 95%), the couples *r*, *n*.

9. References

[Badea 07] A. Badea, O. Gipouloux and J.M. Martinez. Comparaison de métamodèles pour un stockage radioactif. Journées Scientifiques du GdR MoMaS, Frejus, November 14 - 16 2007 (http://www.gdrmomas.org/)

[Bolado 04] R. Bolado and J. Mira. Trivial reductions of dimensionality in the propagation of uncertainties: a physical example. *Environmetrics,* **15**, 57-66, 2004

[Box 87] G.E.P Box and N.R Draper. *Empirical Model Building and Response Surface,* J. Wiley & Sons, Wiley Series in Probability and Mathematical statistics, 1987

[Buckingham 14] E. Buckingham. On physically similar systems. *Physics Review,* **4**, 345 – 376, 1914

[Cannamela 07] C. Cannamela, J. Garnier, and B. Iooss. Controlled stratification for quantile estimation, submitted to *J. Roy. Stat. Soc. B*., 2007

[Conover 80] W.J. Conover. *Practical Nonparametric Statistics*. Second edition. Ed. John Wiley & Sons, 1980

[David 03] H.A. David and H.N. Nagaraja. *Order Statistics, Third ed.,* J. Wiley & Sons, Wiley Series in Probability and Mathematical statistics, 2003

[Doehlert 70] D.H. Doehlert. Uniform shell designs. *Applied Statistics* **19**, 231 – 239, 1970

[Efron 93] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993

[Friedman 91] J. Friedman. Multivariate Adaptative Regression Splines (invited paper), *Annals of Statistics*, Volume **10** (No. 1), 1-141, 1991

[Guba 03] A. Guba, M. Mihakly and P. Lenard. Statistical aspects of best estimate method-1, *Reliability Engineering and System Safety* **80**, 217–232, 2003

[Hastie 90] T. Hastie and R. Tibshirani. *Generalized Additive Models*, Chapman and Hall, London, 1990

[Hastie 02] T. Hastie, R. Tibshirani and J. Friedman. *The elements of statistical learning*. Springer, 2002

[Hickernell 98] F.J. Hickernell. A generalized discrepancy and quadrature error bound. *Math. of Comp*., **67**, 299-322, 1998

[Hyndman 96] R.J. Hyndman and Y. Fan. Sample quantiles in statistical packages, *American Statistician*, **50**, 361–365, 1996

[Iman 82] R.L. Iman and W.J.Conover, A distribution free approach to inducing rank correlations among input variables. *Communications in Statistics, Part B – Simulation and Computation,* **11**, 311-334, 1982

[Jordaan 05] I.J. Jordaan, *Decisions under Uncertainty, Probabilistic Analysis for Engineering decisions*, Cambridge University Press, 2005

[Johnson 90] M.E. Johnson, L.M. Moore and D. Ylvisaker. Minimax and maximin distance designs. *J. of Statis. Planning and Inference*, **26**, 131-148, 1990

[Kooperberg 97] C. Kooperberg, S. Bose and C.J. Stone. Polychotomous Regression, *Journal of the American Statistical Association*, **92**, Issue 437, 117-127, 1997

[Langhaar 51] H.L. Langhaar. *Dimensional Analysis and Theory of Models*, New York, Wiley, 1951

[Lilliefors 67] H. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62, 399-402, 1967

[Makai 06] M. Makai and L. Pal. Best estimate method and safety analysis II**,** *Reliability Engineering and System Safety* **91**, 222–232, 2006

[McKay 79] M.D. McKay, W.J. Conover and R.J. Beckman. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics*, 21, 239-245, 1979

[Mengolini 05] A. Mengolini, K. Simola and R. Bolado. *Formal Expert Judgment. An overview.* EUR 21772 EN, 2005

[Metropolis 49] N. Metropolis and S. Ulam. The Monte Carlo Method*. Journal of the American Statistical association.* Vol. **44**, N. 247, 335-341.

[Mira 04] J. Mira, R. Bolado and P. Solana. The use of dimensional and similarity analysis in the propagation of uncertainties: a physical example. *Journal of computational and Graphical statistics*, **13**, N. 2, 467-486, 2004

[Mood 74] A.M. Mood, F.A. Graybill and D.C. Boes. *Introduction to the theory of statistics*. Third edition. McGraw-Hill International Editions. Statistics series, 1974

[Moran 71] M. Moran. A generalisation of dimensional analysis. *Journal of the Franklin Institute,* **292**, 423-432, 1971

[Moran 72] M. Moran and K.M. Marshek. Some matrix aspects of generalised dimensional analysis. *Journal of Engineering Mathematics,* **6**, 291-303, 1972

[Myers 02] R.H. Myers and D.C. Montgomery. *Response Surface Methodology*, J. Wiley & Sons, Wiley Series in Probability and Statistics, 2002

[Niederreider 87] H. Niederreiter. *Point sets and sequences with small discrepancy*. Monasth. Math., **104**, 273-337, 1987

[NUREG-1150] USNRC ,'Severe accident risks: an assessment of five US nuclear power plants', NUREG-1150, 1990

[Nutt 04] W.T. Nutt and G.B.Wallis. Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties, *Reliability Engineering and System Safety* **83**, 57–77, 2004

[Nutt 05] W.T. Nutt and G.B. Wallis.Reply to "Comments on 'Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties' by W.T. Nutt and G.B. Wallis," by Y. Orechwa, *Reliability Engineering and System Safety* **87**, 137-145, 2005

[Orechwa 05] Y. Orechwa. Comments on 'Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties' by W.T. Nutt and G.B. Wallis, *Reliability Engineering and System Safety* **87**, 133-135, 2005

[Owen 92] A.B. Owen. *Orthogonal arrays for computer experiments, integration and visualiazation.* Statistica Sinica, **2**, 439-452, 1992

[Owen 01] A. B. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, 2001

[Palacios 64] J. Palacios. *Dimensional Analysis.* London, MacMillam, 1964

[Prváková 08] *PAMINA* WP4.3 Benchmark 1 French Clay Repository : Application of Performance Assessment Methodologies for Clay Repository: Uncertainty and Sensitivity Analysis. JRC: S. Prváková, R. Bolado-Lavín, A. Badea and K-F Nilsson; ANDRA: G. Pepin, E. Treille, F. Plas, 2008

[Park 94] J.S. Park. *Optimal latin hypercube designs for computer experiments.* Journal Statistical Planning and Inference, **39**, 95-111, 1994

[Rubinstein 81] R.Y. Rubinstein. *Simulations and Monte-Carlo Method*. Wiley Series in Probability and Mathematical Statistics, J. Wiley & Sons, 1981

[Sacks 89a] J. Sacks, S.B. Schiller and W.J. Welch. Designs for computer experiments, *Technometrics*, 31, pp. 41-47, 1989

[Sacks 89b] J. Sacks, W.J. Welch, T.J. Mitchell and H.P. Wynn, Design and analysis of computer experiments. *Statistical Science*, 4:409-435, 1989

[Santner 03] T.J. Santner, B.J. Williams, N.I. Notz. *The design and analysis of computer experiments*. Springer, 2003

[Saporta 90] G. Saporta. *Probabilités analyses des données et Statistiques*. Editions Technip, Paris, 1990

[Silverman 86] B.W. Silverman. *Density Estimation*. London: Chapman and Hall, 1986

[Simola 05] K. Simola, A. Mengolini, R. Bolado, L. Gandossi. *Training material for formal expert judgment.* EUR 21770 EN., 2005

[Stein 87] M. Stein. Large sample properties of simulations using Latin Hypercube Sampling. *Technometrics*, vol. 29, n°2, 143-151, 1987

[Szirtes 98] T. Szirtes. *Applied dimensional analysis and modeling.* New York, McGraw-Hill, 1998

[Tang 93] B. Tang. *Orthogonal array-based latin hypercubes*. Journal of the American Statistical Association, **88**, 1392-1397, 1993

[Tang 98] B. Tang. *Selecting latin hypercubes using correlation criteria.* Statistica Sinica, **8**, 965-977, 1998

[Thiemard 00] E. Thiémard E. *Sur le calcul et la majoration de la discrépance à l'origine*. Thèse No 2259, Département de mathématiques, école polytechnique fédérale de Lausanne, 2000

[Wallis 03] G.B. Wallis. Contribution to the paper 'Statistical aspects of best estimate method-1' by A. Guba, M. Makai and L. Pal, *Reliability Engineering and System Safety* **80**, 309-311, 2003

[Wallis 06] G.B. Wallis. Evaluating the probability that the outputs of a computer code with random inputs will meet a set of evaluation criteria, *Reliability Engineering and System Safety* **91**, 820-827, 2006

[Welch 92] W.J. Welch, R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell and M. D. Morris. Screening, predicting, and computer experiments, *Technometrics*, 34, 1, 15-25, 1992

[Wilks 41] S. Wilks. Determination of sample sizes for setting tolerance limits; *Ann. Math. Statist*., 12, pp. 91-96, 1941

European Commission

EUR 23503 EN – Joint Research Centre – Institute for Energy
Title: Uncertainty Assessment for PA Models
Author(s): Ricardo BOLADO LAVIN, Anca COSTESCU BADEA
Luxembourg: Office for Official Publications of the European Communities
2008 – 77 pp. – 21x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-09779-9
DOI 10.2790/17625

Abstract
A mathematical model comprises input variables, output variables and equations relating these quantities. The input variables may vary within some ranges, reflecting either our incomplete knowledge about them (epistemic uncertainty) or their intrinsic variability (aleatory uncertainty).

Moreover when solving numerically the equations of the model, numerical errors are also arising. The effects of such errors and variations of the inputs have to be quantified in order to asses the model's range of validity. The goal of uncertainty analysis is to asses the effects of parameter uncertainties on the uncertainties in computed results.

The purpose of this report is to give an overview of the most useful probabilistic and statistic techniques and methods to characterize uncertainty propagation. Some examples of application of these techniques for PA applied to radioactive waste disposal are given.

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LD-NA-23503-EN-C

ISBN 978-92-79-09779-9

Publications Office
Publications.eu.int